世界知的所有権機関
国際事務局

PCT

特許協力条約に基づいて公開された国際出願

AC

(51) 国際特許分類6
H04R 3/02

A1

(11) 国際公開番号 WO98/39946

(43) 国際公開日 1998年9月11日(11.09.98)

(21) 国際出願番号 PCT/JP98/00915

(22) 国際出願日 1998年3月5日(05.03.98)

(30) 優先権データ
特願平9/51577 1997年3月6日(06.03.97) JP

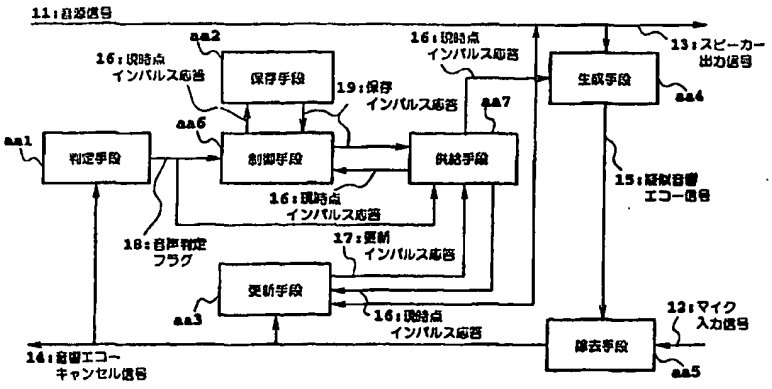
(71) 出願人（米国を除くすべての指定国について）
旭化成工業株式会社
(ASAHI KASEI KOGYO KABUSHIKI KAISHA)[JP/JP]
〒530-8205 大阪府大阪市北区堂島浜1丁目2-6 Osaka, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
庄境 誠(SHOZAKAI, Makoto)[JP/JP]
〒243-0216 神奈川県厚木市宮の里4-1-4-501 Kanagawa, (JP)
谷 智洋(TANI, Tomohiro)[JP/JP]
〒243-0021 神奈川県厚木市岡田4-5-3
厚木ユースハイム115号 Kanagawa, (JP)
(74) 代理人
弁理士 谷 義一(TANI, Yoshikazu)
〒107-0052 東京都港区赤坂5丁目1-31 第6セイコービル3階
Tokyo, (JP)

(81) 指定国 AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO特許 (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), ユーラシア特許 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧州特許 (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI特許 (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

添付公開書類
国際調査報告書

(54)Title: DEVICE AND METHOD FOR PROCESSING SPEECH

(54)発明の名称 音声処理装置および方法



11 ... sound source signal
16 ... current impulse response
aa1 ... judging means
18 ... speech judging flag
14 ... acoustic echo canceling signal
aa2 ... storage means
aa6 ... control means
aa3 ... updating means
19 ... stored impulse response
17 ... updated impulse response
aa7 ... supplying means
13 ... loudspeaker output signal
aa4 ... generating means
15 ... pseudo-acoustic echo signal
12 ... microphone input signal
aa5 ... removing means

(57) Abstract

A speech processor which continuously uses the impulse response used by the preceding frame when a speech signal is included in the microphone input signal or the currently updated impulse response when no speech signal is included in the microphone input signal as the impulse response used for generating a pseudo-acoustic echo signal at the time of performing echo cancellation by using the pseudo-acoustic echo signal.

（５７）要約

擬似音響エコー信号を使用してエコーキャンセルを行う際に、擬似音響エコー信号の発生のために使用するインパルス応答として、マイク入力信号に音声が含まれる場合には前の時点のフレームで使用したインパルス応答を連続的に使用し、マイク入力信号に音声が含まれない場合には新規に更新されたインパルス応答を使用する音声処理装置。

明細書

音声処理装置および方法

技術分野

本発明は、リモートスピーカーからリモートマイクへの音響エコーをキャンセルすることにより通話品質の向上を計ったハンズフリー系の通話システム（テレビ会議システム、自動車電話）およびリモートスピーカーからリモートマイクへの回り込み音声をキャンセルすることにより音声認識性能の向上を計ったハンズフリー系の音声認識装置（カーオーディオ、カーナビゲーション、ＰＣなど）に適用可能な音声処理装置および方法に関する。

背景技術

リモートスピーカーからリモートマイクへ回り込む音響信号は、しばしば音響エコーと呼ばれる。音響エコーを除去する技術（音響エコーキャンセラー）の用途は以下の２つである。

１） ハンズフリー系通話システム（テレビ会議システム、自動車電話）において、通話をしている相手に対し送出される音声の音質を向上させる。

リモートスピーカーから出力された相手側の音声が部屋の壁や窓ガラスに反射し、その部屋固有の音響特性の影響を受けて、リモートマイクに音響的に回り込む場合がある。この場合、相手にとっては自分の声がある時間遅れを伴って音響エコーとして戻ってくるため、聞きづらく話しづらいという不具合がある。従って、リモートマイクが集音した音声の中で、スピーカーから回り込んだ音響エコーをキャンセルして、残りの音声を通話をしている相手に送出することにより、上記の不具合を改善することが望

まれる。

2） ハンズフリー系の音声認識装置において音声認識率を向上させる。

例えば、自動車内においては、カーオーディオやカーナビゲーションのスピーカー出力音が上記音響エコーと同様にダッシュボードや窓ガラスに反射して音声認識用マイクに回り込み、それが非定常の加法性雑音として作用して、音声認識率が低下するという不具合がある。従って、音声認識用マイクが集音した音声の中で、スピーカーから回り込んだ音声をキャンセルして、残りの音声の認識を行うことにより、より高い音声認識性能を実現することが望まれる。

上記２つの用途においてはいずれも、リモートスピーカーからの出力音響の直接音および部屋の壁、ダッシュボードや窓ガラスなどで反射した反射音が常時リモートマイクに回り込む。ここでは、リモートスピーカーからリモートマイクへの直接音および反射音をまとめて音響エコーと呼ぶことにする。また、リモートスピーカー出力音から音響エコーが生成される経路を音響エコー生成経路と呼ぶことにする。

一般に、音響エコー生成経路の特性は、ＦＩＲ(Finite Impulse Response)フィルターでモデル化できるが、部屋内の状況（人間の動作、人数などの要因）や自動車内の状況（人間の動作、人数、窓の開閉などの要因）により変化すると考えられる。音響エコー生成経路の特性の変化がほとんど起こらない場合には、あらかじめ最適なフィルター係数を求めておき、フィルター係数を固定して、音響エコーをキャンセルする方法で良いと思われる。しかしながら、音響エコー生成経路の特性の変化がいつ発生するかは、一般に予測が困難である。この場合、適応フィルターの利用により、最適なフィルター係数を動的に推定しながら、適応的に音響エコーのキャンセルを行う方法の採用が望ましい。

適応フィルターとは、観測信号が、音源既知の信号に対しあるインパルス応答を持つフィルターが畳み込まれて生成されたものであると仮定し、観測信号と（音源既知信号とフィルター係数の推定値との畳込みにより計算される）疑似信号の差が０（ゼロ）になるように、フィルターの係数を動的に適応させるアプローチをいう。音響エコー生成経路を近似するＦＩＲフィルターの係数とスピーカーからの出力信号の畳み込みにより得られる信号を観測信号から引くことにより、音響エコーをキャンセルすることが可能である。適応フィルターのアルゴリズムとして、これまでにＬＭＳ（Least Mean Square error）[S. Haykin, "Adaptive Filter Theory," 2nd ed. Englewood Cliffs, NJ, Prentice-Hall, 1991]、ＮＬＭＳ（Normalized Least Mean Square error）[S. Haykin, "Adaptive Filter Theory," 2nd ed. Englewood Cliffs, NJ, Prentice-Hall, 1991]、ＡＰＡ(Affine Projection Algorithm)[尾関和彦, 南雲仁一, "アフィン部分空間への直交射影を用いた適応フィルター・アルゴリズムとその諸性質," 信学論, Vol.J67-A, No.2, pp.126-132, 1984.]、ＲＬＳ（Recursive Least Squares）[S. Haykin, "Adaptive Filter Theory," 2nd ed. Englewood Cliffs, NJ, Prentice-Hall, 1991]などが提案されている。特に、ＮＬＭＳは、演算量が少なく、収束速度が既知の音源信号の大きさに依存しないため、広く採用されている。しかし、音声のような有色信号に対する、フィルター係数の収束速度が、ＡＰＡやＲＬＳに比べて遅いことが指摘されている。

時刻ｔにおけるＦＩＲフィルターの係数、ＦＩＲフィルターへの入力データ（既知の音源信号）をそれぞれ

$$h(t)=[h_1(t), h_2(t),\ldots,h_M(t)]^T \quad (1)$$

$$x(t)=[x(t), x(t-1),\ldots,x(t-M+1)]^T \quad (2)$$

で表現する。ここで、Tは転置を示す。MはＦＩＲフィルターの次数である。また、時刻tのマイク入力信号をy(t)とすると、ＮＬＭＳは、一般に以下の式で与えられる。

$$r(t)=h(t)^T x(t) \quad (3)$$

$$e(t)=y(t)-r(t) \quad (4)$$

$$h(t+1)= h(t)+\frac{\mu}{a+\|x(t)\|^2}x(t)e(t) \quad (5)$$

ここで、$\|\cdot\|^2$はベクトルのエネルギーを表す。μは、フィルター係数の更新速度を決定する定数（ステップゲインと呼ばれる）で、フィルター係数が収束するために、$0<\mu<2$を満たす必要がある。aは、$\|x(t)\|^2$が微小値の場合に(5)式の右辺第２項が発散するのを防止するための正の定数である。図１に上述の式を回路で表したＮＬＭＳのブロック図を示す。ここで、r(t)を疑似音響エコー信号、e(t)を音響エコーキャンセル信号と呼ぶことにする。図２に、ＮＬＭＳをはじめとする適応フィルターを使用した音響エコーキャンセラー（ＡＥＣ）を室内に設置した例を示す。説明の便宜上、スピーカー２から出力される信号のＡＥＣ１への入力を遠端入力、マイク３の入力を近端入力、スピーカー２の出力を近端出力、音響エコーキャンセル後のＡＥＣ１の出力信号を遠端出力と呼ぶ。また、遠端入力と近端出力は全く等価であると仮定し、遠端入力から近端出力が生成される系の特性（スピーカー特性など）は、音響エコー生成経路の特性に含めるものとする。

このような音響エコーキャンセラーに関しては、特に、以下の課題について精力的に研究が進められてきた。

1）ステップゲインの制御

ステップゲインは可能な限り大きくして収束速度を上げることが必要であるが、大きくしすぎるとハウリングの原因となるため、使用環境に適した設定が必要である。代表的なステップゲインの制御方法として、ＥＳ(Exponential Step)法[S. Makino, Y. Kaneda and N. Koizumi, "Exponentially Weighted Stepsize NLMS Adaptive Filter Based on the Statistics of a Room Impulse Response" IEEE Trans. SAP, Vol.1, No.1, pp.101-108, 1993.]が提案されている。室内における、適応フィルター係数更新時の変化量が指数減衰特性を有することから、ステップゲインを指数的に（変化量の大きいインパルス応答前半では大きく、後半は小さく）設定する。残留エコーレベルが減少するのに要する時間が通常のＮＬＭＳの半分程度で済むことが示されている。

2）ダブルトーク検出

一般に、ダブルトーク（遠端話者と近端話者の双方が同時に発声した状態）において、ＡＥＣ（ＮＬＭＳ）１により適応フィルター係数の更新を継続すると、フィルターの係数が大きく乱れ、その結果としてエコー消去量が減少してハウリングが起こりやすくなる。従って、ダブルトークを如何に速やかに検出して、ＡＥＣ１の適応フィルターの更新動作を制御（停止・再開）するかは重要なポイントである。ダブルトークの検出には、残留エコーのパワーを用いる方法が有効とされる[藤井健作, 大賀寿郎, "エコー経路変動検出を併用するダブルトーク検出法," 信学論, Vol.J78-A, No.3, pp.314-322, 1995.]。音響エコーに埋もれる小さな、近端話者音声を検出することが可能であるからである。

3）音響エコー経路変動検出

近端話者が移動した時などに伴う音響エコー経路の変動が発生した場合、

残留エコーが増大してダブルトークと判断され、適応フィルターの係数更新が停止されるという不具合が生じる。そこで、ダブルトークと音響エコー経路変動を区別し、音響エコー経路変動の場合には適応フィルターの係数更新を継続することが必要になる。その方法として、近端入力と疑似エコーの相関を利用する方法[藤井健作, 大賀寿郎, "エコー経路変動検出を併用するダブルトーク検出法," 信学論, Vol.J78-A, No.3, pp.314-322, 1995.]などが提案されている。

カーオーディオやカーナビの音響・音声信号がスピーカーから出力されている自動車環境内でのロバストな音声認識機能を高める目的で、車室内にＮＬＭＳによるＡＥＣ１を設置して音源既知の加法性雑音の除去する例を図３に示す。図３に示す図２と同一個所の符号は図２に示す符号と同一である。

スピーカー２からガイダンス音声が出力されている最中に発声された音声を認識できる機能、いわゆるBarge-In(Talk-Through)機能の実現に向けて、ＡＥＣ１を利用する手法が試みられている。ここで、スピーカー２の出力に起因する音声の誤認識の中で、適応フィルターの効果により正認識となる回復率をＲＲＥ(Recovery Rate of Error)と呼ぶことにする。

例えば、会議室に設置された音声対話システムにおいて、スピーカー２からマイク３へのガイダンス音声の回り込みをＡＥＣ１により抑圧することにより、７０～８０％のＲＲＥが得られることが示されている[高橋敏, 嵯峨山茂樹, "ＮＯＶＯ合成法を用いたＢａｒｇｅ－ｉｎ音声の認識," 音講論集, 2-5-1, pp.59-60, 1996-3.]。

しかしながら、音源未知の加法性雑音が存在し、その雑音レベルが常時変動する車室内における、音響エコーキャンセラーに関する研究成果はあまり報告されていない。自動車電話のハンズフリー装置においては、音声

スイッチ（近端入力と遠端入力のエネルギー比較による交互通話方式）と音響エコーキャンセラーとの併用によるものもあるが、語頭、語尾の切断が多く通話品質が不十分である点が指摘されている。

一般に、近端入力に近端出力から生成される音以外の音が混入する（以下、近端入力が存在するという）状況で係数の適応化を継続した場合、フィルター係数の推定精度が劣化し、音響エコーのキャンセル性能が悪化する。そこで、遠端入力が存在し、かつ近端入力が存在する状態（ダブルトーク状態と呼ぶ）では、(5)式によるフィルター係数の更新を停止させることが一般に行われる。遠端入力が存在するかどうかの判断は、遠端入力のエネルギーと予め定められたしきい値との単純な比較で可能である。

一方、近端入力が存在するかどうかの判断を同様に行うと、音響エコーの影響で近端入力が存在すると判断する場合が多くなり、(5)式によるフィルター係数の更新を頻繁に停止して、結果的にフィルター係数の推定精度が劣化するという不具合が生じる。そこで、近端入力信号 y(t)ではなく、音響エコーキャンセル信号 e(t)のエネルギーを用いて、近端入力が存在するかどうかを判断するという方法が考えられる。近端出力から生成された音以外で近端入力に混入する音としては、大きく分けて走行雑音などの音源未知の加法性雑音と人間の音声の２つが考えられるが、いずれも適応フィルターで除去されずに遠端出力に残存する。

一般に、走行中の自動車環境では、音源未知の加法性雑音のエネルギーレベルが、６０～８０ｄＢＡの間で大きく変動するため[金指久則, 則松武志, 新居康彦, "車載用単語音声認識装置," 音講論集, 1-Q-32, pp.159-160, 1995-3.][鈴木邦一, 中村一雄, 宇尾野豊, 浅田博重, "車載騒音環境下における連続音声認識," 音講論集, 2-Q-4, pp.155-156, 1993-10.]、近端入力の存在を判断するための最適なしきい値を一意に決めることは難しいという問題

がある。

また、音源未知の加法性雑音の影響により、近端入力と疑似音響エコー信号の相関が低下するため、先述の音響エコー経路変動検出法[藤井健作, 大賀寿郎, "エコー経路変動検出を併用するダブルトーク検出法," 信学論, Vol.J78-A, No.3, pp.314-322, 1995.]の適用も困難な場合があると予想される。音源未知の加法性雑音と人間の音声を正確に識別する能力を持つ、音声検出アルゴリズムがあれば、有力な解決法になると思われる。

まず走行雑音のみが存在する場合での、ＮＬＭＳによる音響エコーのキャンセル性能を評価する。図４Ａ、図４Ｂ、図４Ｃ、図４Ｄ、図４Ｅに、それぞれ遠端入力信号（ポップス音楽）のスペクトログラム、アイドリング時での近端入力信号のスペクトログラム、同じく音響エコーキャンセル信号のスペクトログラム、時速１００ｋｍ走行時の近端入力信号のスペクトログラム、同じく音響エコーキャンセル信号のスペクトログラムを示す。

カーオーディオの音量は、アイドリング時と時速１００ｋｍ走行時で、男性１名が快適と感じるレベルにセットした。従って、時速１００ｋｍ走行時の方が、スピーカー出力レベルは大きく、音響エコーレベルも大きい。近端入力信号は、２０００ｃｃの自動車の運転席サンバイザーに単一指向性マイクを設置して収録した。フィルター係数の初期値は全て０．０とし、時刻０秒から継続的に(3)-(5)式によりフィルター係数を更新しながら音響エコーキャンセル信号を求めた。サンプリング周波数は８ｋＨｚであり、音響エコーの最大遅延は３２ｍｓまで考慮した。従って、ＦＩＲフィルターのタップ数は２５６である。

また、適応フィルターの性能を評価する尺度として、ＥＲＬＥ(Echo Return Loss Enhancement)がよく用いられる。ＥＲＬＥは近端入力信号の減衰量を表し、次式で定義される[北脇信彦編著, "音のコミュニケーション工

学ーマルチメディア時代の音声・音響技術ー," コロナ社, 1996.]。

$$ERLE=10\cdot\log_{10}\frac{E[y(t)^2]}{E[e(t)^2]} \qquad (6)$$

E[·]は推定値を表し、次式により求める。

$$E[z(t)^2]=(1-\lambda)\cdot E[z(t-1)^2]+\lambda\cdot z(t)^2 \qquad (7)$$

但し、λ＝１／２５６である。ＥＲＬＥの単位は、ｄＢである。アイドリング時のＥＲＬＥの最大値、平均値はそれぞれ１８．８０ｄＢ、１０．１３ｄＢである。また、時速１００ｋｍ走行時のＥＲＬＥの最大値、平均値はそれぞれ９．３３ｄＢ、５．８９ｄＢである。近端入力の音源未知の加法性雑音のレベルが大きいほど、(6)式で与えられるＥＲＬＥは低い値になることに注意する必要がある。

図４Ｃ、図４Ｅからアイドリング時、時速１００ｋｍ時いずれの場合も音響エコーをほぼキャンセルできていることが分かる。近端入力に人間の音声が含まれない場合は、フィルター係数を継続的に更新することにより音響エコーの大部分はキャンセル可能であると思われる。すなわち、音源未知の加法性雑音の中で定常的かつ音声と無相関である走行雑音は、フィルター係数の推定に与える影響が小さいと考えられる。

次に、近端入力に人間の音声が含まれる場合について調べる。２０００ｃｃの自動車でカーオーディオからポップス音楽を再生しながら市街地を時速６０ｋｍで走行し、加法性雑音データを収録した。この時、音楽のボリュームは女性１名が快適と感じるレベルにセットした。次に、停止中（エンジンオフ）の同一の自動車内で同一女性１名が発声した音声データ（「明るい」）を同一の録音レベルで収録した。

そして、加法性雑音データと音声データとを計算機上で加算した信号のスペクトログラムを図７Ａに示す。図７Ｂにフィルター係数の初期値を０．０とし、時刻０秒から連続的にフィルター係数を更新した場合の音響エコーキャンセル信号のスペクトログラムを示す。また、図７Ｃにフィルター係数の１０番目の係数の値の変化を示す。この時の、ＥＲＬＥの最大値、平均値はそれぞれ８．４８ｄＢ、４．１８ｄＢである。

特に、時刻０．５秒あたりから０．１５秒の間、フィルター係数値が激しく振動し、不安定になっている様子が分かる。また、時刻１．０秒以降の音響エコー（図７Ｂの楕円で囲まれた部分）をキャンセルできていない。近端入力に音声が存在する間はフィルター係数の更新を停止し、近端入力に音声が存在しない間は、定常的な加法性雑音の存在の如何に関わらずフィルター係数の更新を継続する必要がある。そのためには、音源未知の加法性雑音が混入する近端入力に音声が含まれているかどうかを正確に判定する音声検出アルゴリズムが必要となる。

音声認識システムにおいては、正確に音声区間を検出すること（音声検出）が極めて重要である。背景雑音がほとんどない環境では、正確な音声検出はそれほど難しくはない。しかしながら、走行中の車室内のように背景雑音の存在が無視できない環境においては、音声の検出はかなり困難である。特に、音声の最初に位置する弱い摩擦音、弱い鼻音や音声の最初または最後に位置する無声化した母音などは背景雑音に埋もれてしまうケースが多く、検出は難しい。呼吸音、舌打ち音などは本来非音声として検出されるべきであるが、しばしば音声として検出され、誤認識につながることが多い。

通常、あるしきい値以上の短時間パワーが連続して一定フレーム以上継続するかどうかにより音声の開始点を検出し、あるしきい値以下の短時間

パワーが連続して一定フレーム以上継続するかどうかにより音声の終了点を検出する方法が一般的である。また、２つのレベルのしきい値を用いて、より正確に音声を検出しようとする試みや、音声信号の零交差回数を用いることもある[古井貞熙, "ディジタル音声処理," ディジタルテクノロジーシリーズ, 東海大学出版会, 1985.]。音源未知の加法性雑音の存在が無視できる環境においては、短時間パワーや零交差回数などの時間情報のみを用いる音声検出法でも問題は生じない。

しかし、音源未知の加法性雑音の存在が無視できない環境においては、従来の音声検出法を用いた場合の音響エコーキャンセラーには、以下の不具合がある。まず、第一にマイク入力に音声が存在しないにも関わらず、音源未知の加法性雑音を音声であると判断し、フィルター係数の更新が行われず、音響特性の変化に追随できなくなり、音響エコーのキャンセル性能が低下する。第２に、マイク入力に音声が存在するにも関わらず、音声がないと判断し、フィルター係数の更新が行われて、所望の値から乖離し、音響エコーのキャンセル性能が低下する。従って、時間情報ばかりではなく、スペクトルなどの周波数情報も併用する方法が望ましい。

特開平９－２１３９４６号（ＮＴＴ）においては、入力音声信号（エコーキャンセル前の信号）と音源既知の加法性雑音の音源情報の時間情報および周波数情報を用いて入力音声信号に音声が含まれているかどうかを判定するダブルトーク検出回路を用いた音響エコーキャンセラーが説明されている。しかしながら入力音声信号に入り込むエコーとして音源信号の影響を受けたものだけを前提にしており、周囲の雑音がある場合に、ダブルトーク検出精度が悪いという不具合がある。また、適応フィルターにより推定したインパルス応答（ＦＩＲフィルターの係数値）を保持するバッファを有していない。

特開平５－１０２８８７号（東芝）では、エコーキャンセル後の信号の大きさでダブルトークかどうか判定するダブルトーク検出回路を用いているが、時間情報および周波数情報を併用する判定ではないため、周囲の雑音が存在する環境における判定精度が十分ではないという不具合がある。

特開平７－３０３０６６号（ＮＴＴ　ＤＯＣＯＭＯ）では、インパルス応答レジスタで判定手段の遅れを補償する構成を取っているが、エコーキャンセル後の信号の時間情報および周波数情報を用いて入力音声信号に音声が含まれているかどうかをフレーム毎に判定する手段を具えていないため、ダブルトーク検出性能に限界がある。

ＷＯ　９６／４２１４２号（ＮＯＫＩＡ）では、エコーキャンセル後の信号の時間情報および周波数情報を用いて入力音声信号に音声が含まれているかどうかをフレーム毎に判定する手段を具えているが、自動車電話の基地局の送出信号のゲインを小さくすることにより音響エコーが直接送出されるのを押さえる構造を持つ音響エコーサプレッサに関する発明であり、音響エコーキャンセラーに関する発明ではない。

発明の開示

本発明の目的は、音響などの雑音が混在しやすい環境下で音声信号からの雑音除去性能を改善することができる音声処理装置および方法を提供することにある。

音響エコーの伝達経路を模擬する現時点のインパルス応答および音源信号に基づき疑似音響エコー信号を生成する生成手段と、

現時点のインパルス応答を保持し、前記生成手段に供給する供給手段と、

マイク入力信号から該疑似音響エコー信号を減算することにより音響

エコー成分を除去し、音響エコーキャンセル信号を生成する除去手段と、

前記音源信号と前記音響エコーキャンセル信号と前記供給手段が保持している現時点のインパルス応答を用いて継続的にインパルス応答を更新し、更新されたインパルス応答を前記供給手段に供給する更新手段と、

前記音響エコーキャンセル信号の時間情報および周波数情報を利用して、マイク入力信号に音声が含まれているか否かをフレーム毎に判定する判定手段と、

１つ以上のインパルス応答を保存する保存手段と、

前記判定手段の判定結果が否定判定のフレームでは前記供給手段が保持している現時点のインパルス応答を前記保存手段に保存し、肯定判定のフレームでは前記保存手段に保存されているインパルス応答の１つを取り出して、前記供給手段に供給する制御手段と

を具えたことを特徴とする。

本発明では前記音響エコーがキャンセルされた後の信号を音声認識に用いてもよい。

本発明ではさらに前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルに基づき各フレーム毎に連続的にスペクトル平均を求める手段と、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信号から各フレーム毎に計算されたスペクトルから連続的に減算することにより、音源未知の加法性雑音を除去する手段とを具えてもよい。

本発明ではさらに前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルから各フレーム毎に連続的にスペクトル平均を求める手段と、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信

号から各フレーム毎に計算されたスペクトルから連続的に減算することにより、音源未知の加法性雑音を除去する手段と、当該加法性雑音が除去されたスペクトルからケプストラムを求める手段と、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求め手段と、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算して、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正する手段とを具えてもよい。

本発明ではさらに、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルからケプストラムを求める手段と、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求める手段と、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算することにより、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正する手段とを具えてもよい。

本発明では、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルからからケプストラムを求める手段と、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求める手段と、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算することにより、マイク特性や口からマイ

クまでの空間伝達特性に依存する乗法性歪みを補正する手段とを具えてもよい。

本発明では、擬似音響エコー信号を使用してエコーキャンセルを行う際に、擬似音響エコー信号の発生のために使用するインパルス応答として、マイク入力信号が音声の場合には前の時点のフレームで使用したインパルス応答を連続的に使用し、マイク入力信号が音声ではない場合には新規に更新されたインパルス応答を使用することで音響エコーキャンセリングの性能を改善する。

さらに本発明は音響エコーをキャンセルした後の信号からフレーム毎のスペクトルおよびスペクトル平均を求め、得られたスペクトルおよびスペクトル平均を使用して加法性雑音を除去する。

図面の簡単な説明

図１はＮＬＭＳ(Normalized Least Mean Square error)の機能構成を示すブロック図である。

図２は音響エコーキャンセラーの設置例を示す図である。

図３は車室内における音源既知の加法性雑音を除去する例を示す図である。

図４Ａ－図４ＥはそれぞれＮＬＭＳ(Normalized Least Mean Square error)の性能（横軸：ｓｅｃ．）を示す図である。

図５はＶＡＤ(Voice Activity Detection)の処理内容を示すブロック図である。

図６はＶＡＤの動作タイミングを示す図である。

図７Ａ－図７ＧはそれぞれＮＬＭＳ－ＶＡＤ(Normalized Least Mean Square error with frame-wise Voice Activity Detection)の効果（横軸：ｓｅ

ｃ．）を示す図である。

図８はフィルター係数バッファの動作を説明するための図である。

図９はＮＬＭＳ－ＶＡＤの構成を示すブロック図である。

図１０Ａおよび図１０Ｂはそれぞれ ＮＬＭＳ－ＶＡＤ／ＣＳＳ法によるスペクトログラムを示す図である（横軸：ｓｅｃ．）。

図１１は時不変フィルタを示す図である。

図１２はＮＬＭＳ－ＶＡＤ／ＣＳＳ／Ｅ－ＣＭＮ法の処理内容を示すブロック図である（横軸：ｓｅｃ．）。

図１３はＮＬＭＳ－ＶＡＤ／ＣＳＳ／Ｅ－ＣＭＮの評価を示す図である。

図１４は本発明第１実施形態の音声処理装置の構成を示すブロック図である。

図１５は本発明第２実施形態のシステムの構成を示すブロック図である。

図１６は本発明第３実施形態のシステムの構成を示すブロック図である。

図１７は本発明第４実施形態のシステムの構成を示すブロック図である。

図１８は本発明第５実施形態のシステムの構成を示すブロック図である。

図１９は本発明第６実施形態のシステムの構成を示すブロック図である。

発明を実施するための最良の形態

短時間パワーやピッチなどの時間情報とスペクトルなどの周波数情報を利用した音声検出アルゴリズムの１つとして、欧州の携帯電話・自動車電話システムであるＧＳＭで標準規格化されている音声検出ＶＡＤ(Voice Activity Detection)[Recommendation GSM 06.32.]がある。このＶＡＤは音声ＣＯＤＥＣ（圧縮・伸張）などのデジタル信号処理の動作を細かく制御し、低消費電力化を計って電池寿命を延ばす目的で用いられている。図５にこのＶＡＤの簡単な構成を示す。まず、音声信号からフレーム毎に自己

相関関数（時間情報）が求められる。この自己相関関数から線形予測分析ＬＰＣ（Linear Predictive Coding）により、線形予測係数（時間情報）が求められる。線形予測係数から構成できる逆ＬＰＣフィルターと自己相関関数から音声信号の短時間パワー（時間情報）を求めることができる。この短時間パワーとしきい値を比較し、ＶＡＤ判定を行う。

短時間パワーがしきい値よりも大きい場合は、値１の局所的なＶＡＤフラグが出力される。そうでない場合は、値０（ゼロ）の局所的なＶＡＤフラグが出力される。そして、ＶＡＤ後処理において過去の複数のフレームの局所的なＶＡＤフラグの値の履歴を用いて最終的なＶＡＤフラグの値が決定される。

一方、ＶＡＤ判定において短時間パワーとの比較に用いられるしきい値は、以下のように適応化される。平滑化された自己相関関数と自己相関予測係数により表されるスペクトル（周波数情報）変化が連続するフレーム間で十分小さい場合は、スペクトルの定常性が高いと判断される。スペクトルの定常性が高いと判断される音声信号としては背景雑音または母音が考えられる。

背景雑音のフレームにおいてはしきい値適応を行い、母音のフレームではしきい値適応を行うべきではない。背景雑音と母音を区別するためにピッチ情報を利用する。音声信号から計算される自己相関関数からピッチラグ（ピッチ周期）（時間情報）が計算される。連続するフレーム間でピッチラグの変化が小さい場合は、そのフレームは母音であると判断され、値１のピッチフラグが出力される。そうでない場合は、値０（ゼロ）のピッチフラグが出力される。

上記の短時間パワー、逆ＬＰＣフィルターから求められる残差信号自己相関予測係数、ピッチフラグ、定常性の情報を利用してスペクトルの定常

性が高く、ピッチ性が低いフレームにおいてしきい値の適応化が行われる。このＶＡＤはエンジン音やロードノイズなどの比較的定常的な背景雑音にたいしてはそのレベルに関わらず、正確な音声検出性能を発揮する。

自動車内において、安全性の見地からマイクがサンバイザーなど口元から離れた場所に設置される場合、信号雑音比（ＳＮＲ）は１０ｄＢ以下にまで悪化する。その場合には、上記のＶＡＤアルゴリズムの音声検出性能は著しく劣化することが分かった。

そこで、ＳＮＲ１０ｄＢ程度でも正確に音声検出ができるようにしきい値の適応化などの部分を改良した。現在、ＶＡＤに用いている音声の窓長は３２ｍｓ、フレームシフトは１０ｍｓである。以後、ＶＡＤにより音声の存在が検出された場合、ＶＡＤがＯＮであると言う。逆に、検出されなかった場合、ＶＡＤがＯＦＦであると言う。このＶＡＤは、１フレームに１回近端入力に音声が含まれているかどうかの判断を行うため、音声の検出タイミングが実際の音声の立ち上がりから遅延することが起こりえる。

図６に、フレーム、ＶＡＤの動作タイミングおよびＶＡＤが使用する窓長の関係を示す。実際の音声の開始がフレームｎの中心である場合、ＶＡＤによりその音声の開始を検出できるのは、フレームｎ＋１以降である可能性が高い。仮に、フレームｎ＋２のＶＡＤで検出できた場合、実際の音声の開始からの検出遅れは２５ｍｓにもなり、その間、エコーパスの推定値が不安定になることが考えられる。

図７ＤにＶＡＤによる音声検出の様子を示す。レベル１が音声が検出されたことを示す。矢印を用いて示すように、２フレーム程度の音声検出遅れが認められる。不安定になったフィルター係数値をより精度の高い値に回復することができれば、音響エコーキャンセル性能の低下を避けることが可能だと考えられる。

そこで、m個分のフィルター係数を格納できるバッファ（フィルター係数バッファと呼ぶ）を用意する。ＶＡＤがＯＦＦのフレームでは、ｎ（ｍ－１≧ｎ≧１）番目の格納位置に格納されたフィルター係数を順次ｎ＋１番目の格納位置に移すと同時に、現時点での適応フィルターの係数をフィルター係数バッファの第１番目の格納位置に格納する。この時、結果として、m番目の格納位置に格納されていたフィルター係数は捨てられる。一方、ＶＡＤがＯＮのフレームでは、フィルター係数バッファのm番目の格納位置に格納されたフィルター係数を取り出し、その値で劣化したフィルター係数をリセットすれば良い。

図８にフィルター係数バッファの動作の様子を示す。ｍを０～４とした場合のＥＲＬＥの最大値、平均値を表１に示す。

表１　フィルターバッファのサイズとＥＲＬＥ(Echo Return Loss Enhancement)の関係

buffer size m	max ERLE(dB)	average ERLE(dB)
0	8.80	4.18
1	9.06	4.25
2	9.15	4.35
3	9.14	4.36
4	9.14	4.36

m＝０は係数値の保存およびリセットを行わない場合を示す。m≧２では、ＥＲＬＥにほとんど違いが見られないため、m＝２を選択する。これは、ＶＡＤの検出遅れ（２フレーム程度）と対応している。

上記の特徴を持ったアルゴリズムをＮＬＭＳ－ＶＡＤ（NLMS with frame-wise VAD）と呼び、全体のブロック図を図９に示す。ここで、[s]、[f]はそれぞれサンプルワイズ、フレームワイズの信号の流れおよび処理の動作を示す。ＶＡＤがいったんＯＮになると、次にＶＡＤがＯＦＦになるフレームまでフィルター係数の更新は停止される。図７Ｅに、フィルター係数の初期値を全て０．０とし、ＶＡＤを動作させ、フィルター係数値の格納およびリセットを行いながら、時刻０秒からフィルター係数を更新した場合の音響エコーキャンセル信号のスペクトログラムを示す。図７Ｆにその時のフィルター係数の１０番目の係数の値の変化を示す。フィルター係数の更新が停止されたフレームの直前で、フィルター係数値が不安定になっているが、上記フィルター係数の格納およびリセットにより、フィルター係数が回復されている様子が示されている。これにより、時刻１．０秒以降の音響エコー（図７Ｂの楕円で囲まれた部分）もキャンセルされている。

但し、図７Ｅで時刻０．１秒前後の音響エコー（図７Ｅの楕円で囲まれた部分）がキャンセルされていないことが分かる。発声毎に推定されたフィルター係数およびＶＡＤに用いられるパラメータを保存しておき、次の発声時にそれらを初期値として用いれば、フィルター係数の推定速度は速まると考えられる。図７Ｇにその例を示す。時刻０．０秒直後の音響エコーは若干残存しているが、それ以後の音響エコー（図７Ｅの楕円で囲まれた部分）はほぼキャンセルされていることが分かる。この時の、ＥＲＬＥの最大値、平均値はそれぞれ９．２９ｄＢ、４．５０ｄＢである。また、本

願出願人はＮＬＭＳ－ＶＡＤ法に関連して、時間情報および周波数情報に基づく音声検出を用いた音響エコーキャンセラーの日本国出願を既に済ませている（特願平０９－０５１５７７号、１９９７年３月６日出願）。なお、本願発明では時間情報および周波数情報に基づく音声検出をフレーム毎に行う点が、上記先願発明との相違点である。

次に、音源既知の加法性雑音および音源未知の加法性雑音が存在する環境におけるロバストな音声認識方法として、ＮＬＭＳ－ＶＡＤ法とＣＳＳ(Continuous Spectral Subtraction)法を組み合わせる方法について説明する。時刻ｔにおける周波数ωでの観測スペクトル、音声スペクトル$S(\omega;t)$の推定値、加法性雑音の推定値をそれぞれ$O(\omega;t)$、$\hat{S}(\omega;t)$、$\hat{N}(\omega;t)$と表すとすると、ＣＳＳ法は以下のように与えられる。

$$\hat{N}(\omega;t)=\gamma\cdot\hat{N}(\omega;t-1)+(1-\gamma)\cdot O(\omega;t) \tag{8}$$

$$\hat{S}(\omega;t)=\begin{cases}O(\omega;t)-\alpha\cdot\hat{N}(\omega;t) & \text{if } O(\omega;t)-\alpha\cdot\hat{N}(\omega;t)>\beta\cdot O(\omega;t)\\ \beta\cdot O(\omega;t) & \text{otherwise}\end{cases} \tag{9}$$

ここで、αはover-estimation factor、βはflooring factor、γはsmoothing factorであり、以下では予備実験の結果から、それぞれ２．４、０．１、０．９７４と設定した。ＣＳＳは、音声フレームと非音声フレームを区別せず、連続的にスペクトルの移動平均を求め、これを雑音スペクトルの推定値とみなして、入力スペクトルから減算する方法である。雑音スペクトルの推定値に音声スペクトルの影響が含まれるため、エネルギーの弱い音声スペクトルがマスクされてしまい、歪みが生じるという問題点があるが、過去のある一定時間長の区間に対して、相対的に大きなエネルギーを持つ周波

数成分を残し、エネルギーの微弱な周波数成分を雑音、音声を問わず、マスクするという働きを持つ。このため、クリーンな音声にＣＳＳを施した後に得られる特徴パラメータと加法性雑音が重畳した音声にＣＳＳを施した後に得られる特徴パラメータの間の変動が、通常のスペクトル減算法や最小平均二乗誤差推定法に比べて小さい。この特長は、低いＳＮＲでの音声認識にとって有効である。図１０Ａに、停止中（アイドリング）の自動車内で女性が発声した音声（「明るい」、図７Ａに示した音声を計算機上で加算して作成した際に用いた音声と同一）にＣＳＳを施した後のスペクトログラムを、図７Ｂに同一音声に時速６０ｋｍ走行時の音源未知の加法性雑音と音響エコーが重畳した雑音データを計算機上で加算した後（図７Ａ）、ＮＬＭＳ－ＶＡＤ法で音響エコーをキャンセルし（図７Ｇ）、ＣＳＳ法を施して得られるスペクトログラムを示す。図７Ｇと図１０Ｂを比較すると、時刻０．９秒近辺の周波数１ｋＨｚの音響エコーの残存成分（図７Ｇの楕円で囲まれた部分）がＣＳＳ法により除去されていることが分かる。

ＣＳＳ法は、定常的な加法性雑音だけでなく、ＮＬＭＳ－ＶＡＤ法でキャンセルできなかった残存音響エコーを抑圧する効果も持っている。音響エコーキャンセル信号e(t)にＦＦＴを施して得られたスペクトルに対してＣＳＳ法を施した後のスペクトルを逆ＦＦＴにより時間領域に戻して得られる波形信号を、(6)式のe(t)の代わりに用いた場合のＥＲＬＥの平均値は１３．６０ｄＢであった。これに対し、ＮＬＭＳ－ＶＡＤ法による音響エコーキャンセルを行わず、ＣＳＳ法による加法性雑音のキャンセルのみを行って同様に求めたＥＲＬＥの平均値は９．８７ｄＢであった。ＣＳＳのみでは、約３．７ｄＢ相当の音源既知の加法性雑音がキャンセルできなかったと見ることができる。

図１０Ａと図１０Ｂを比較すると２つのスペクトログラムがきわめて類似していることが分かる。ＮＬＭＳ－ＶＡＤ法とＣＳＳ法の組み合わせにより、音源既知の加法性雑音と音源未知の加法性雑音に対して、ロバストな特徴パラメータを抽出できることが示唆されている。

次に、音声スペクトルに対する乗法性歪みの補正方法について述べる。ある個人の発声器官で生成される、時刻ｔにおける周波数ωでの短時間スペクトル$S(\omega;t)$の音声フレームにおける長時間平均を話者の個人性$H_{Person}(\omega)$と呼ぶこととし、

$$H_{person}(\omega)=\frac{1}{T}\cdot\sum_{t=1}^{T}S(\omega;t) \qquad (10)$$

と定義する。ここで、Ｔは十分大きな自然数である。$H_{Person}(\omega)$は、声帯音源特性および声道長に依存する話者固有の周波数特性を表しているとみなすことができる。また、短時間スペクトルを話者の個人性で除したもの

$$S^{*}(\omega;t)=S(\omega;t)/H_{Person}(\omega) \qquad (11)$$

を正規化音声スペクトルと定義する。この時、図１１に示すように、音声スペクトルは、正規化音声スペクトル$S^{*}(\omega;t)$が時不変フィルター$H_{Person}(\omega)$を通過することにより生成される、あるいは、正規化音声スペクトル$S^{*}(\omega;t)$に乗法性歪み$H_{Person}(\omega)$が重畳して生成されると解釈することができる。

$$S(\omega;t)=H_{Person}(\omega)\cdot S^{*}(\omega;t) \qquad (12)$$

車室内のような実環境においては、正規化音声スペクトルに対する乗法性歪みとして、上記の話者の個人性に加えて、以下の３種類が考えられる[A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Kluwer Academic Publishers, 1992.]。

（１）発話様式 $H_{Style(N)}(\omega)$

加法性雑音Ｎに依存する発話様式（しゃべり方、発話速度，発話の大きさ，Lombard効果など）に固有の周波数伝達特性である。Lombard効果とは、加法性雑音が存在する環境下で発声をする場合に、静寂な環境下とは異なって、無意識のうちに発声スペクトルが変形する現象のことをいう。文献[Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," IEEE Trans. ASSP, Vol.36, No.4, pp.433-439, 1988.]では、ソフトなしゃべり方の場合は、１ｋＨｚ以下のエネルギーが強く、１ｋＨｚ以上のエネルギーが弱いという特性があること、一方で、大きな声、早口、叫び声、Lombard効果の場合は、逆の特性を持つことが示されている。

（２）空間伝達特性 $H_{Trans}(\omega)$

口からマイクまでの空間的な周波数伝達特性を表す。

（３）マイク特性 $H_{Mic}(\omega)$

マイクなどの入力系の電気的な周波数伝達特性を表す。

一般に，音声と雑音の線形スペクトル領域での加法性が成り立つとすると、時刻ｔにおける周波数ωでの観測スペクトル$O(\omega;t)$は、

$$O(\omega;t) = H_{Mic}(\omega)\cdot[\, H_{Trans}(\omega)\cdot\{H_{Style(N)}(\omega)\cdot(H_{Person}(\omega)\cdot S^{*}(\omega;t))\} + N(\omega;t) + E(\omega;t)] \quad (13)$$

でモデル化できる[J. H. L. Hansen, B. D. Womack, and L. M. Arslan, "A Source Generator Based Production Model for Environmental Robustness in Speech Recognition," Proc. ICSLP 94, Yokohama, Japan, pp.1003-1006, 1994.]。ここで、$N(\omega;t)$は音源未知の加法性雑音スペクトルを表し、$E(\omega;t)$は音源既知の加法性雑音スペクトルを表す。

４種類の乗法性歪みの内、$H_{Mic}(\omega)$はあらかじめ測定可能であるが、$H_{Person}(\omega)$、$H_{Style(N)}(\omega)$、$H_{Trans}(\omega)$を，実環境において音声認識システムのユ

ーザーに負荷をかけることなく分離して測定することは困難であると考えられる。また、例え加法性雑音$N(\omega;t)$、$E(\omega;t)$が存在しないとしても、観測スペクトルの長時間平均として(10)式と同様に求められた時不変フィルターのゲインには、上記4種類の乗法性歪みの混在が避けられない。そこで、改めて乗法性歪み$H^*(\omega)$、加法性雑音$\tilde{N}(\omega;t)$、$\tilde{E}(\omega;t)$を、それぞれ

$$H^*(\omega) = H_{Mic}(\omega) \cdot H_{Trans}(\omega) \cdot H_{Style(N)}(\omega) \cdot H_{Person}(\omega) \quad (14)$$

$$\tilde{N}(\omega;t) = H_{Mic}(\omega) \cdot N(\omega;t) \quad (15)$$

$$\tilde{E}(\omega;t) = H_{Mic}(\omega) \cdot E(\omega;t) \quad (16)$$

と定義すると、(13)式を以下のように簡単化できる。

$$O(\omega;t) = H^*(\omega) \cdot S^*(\omega;t) + \tilde{N}(\omega;t) + \tilde{E}(\omega;t) \quad (17)$$

一方、(17)式を変形すると、

$$S^*(\omega;t) = \frac{O(\omega;t) - \tilde{N}(\omega;t) - \tilde{E}(\omega;t)}{H^*(\omega)} \quad (18)$$

が得られる。不特定話者音素モデルを観測されたスペクトルではなく、(11)式により正規化されたスペクトルを用いて作成しておけば、観測スペクトル$O(\omega;t)$に対し、実環境における$\tilde{N}(\omega;t)$、$\tilde{E}(\omega;t)$、$H^*(\omega)$の除去を行って、正規化音声スペクトル$S^*(\omega;t)$の推定値を求めることにより、頑健な音声認識システムを実現できると考えられる。$\tilde{N}(\omega;t)$、$\tilde{E}(\omega;t)$の除去に関して

は、ＮＬＭＳ－ＶＡＤ法とＣＳＳ法を組み合わせる方法が有効であることを先述した。

音声認識システムにおいては、音響パラメータとして、通常、スペクトルの代わりに、ケプストラムが用いられる。ケプストラムは、スペクトルの対数値に逆離散コサイン変換（ＤＣＴ：Discrete Cosine Transform）を施したものとして定義される。ケプストラムはスペクトルに比べて、少ないパラメータ数で同等の音声認識性能が得られるためよく用いられる。

正規化音声スペクトル$S^*(\omega;t)$に対する乗法性歪み$H^*(\omega)$の除去に関しては、次のＥ－ＣＭＮ(Exact Cepstrum Mean Normalization)法が有効であることが既に示されている[M. Shozakai, S. Nakamura and K. Shikano, "A Non-Iterative Model-Adaptive E-CMN/PMC Approach for Speech Recognition in Car Environments," Proc. Eurospeech, Rhodes, Greece, pp.287-290, 1997.]。Ｅ－ＣＭＮ法は次の２つのステップから構成される。

推定ステップ：話者毎に音声／非音声フレームで別々にケプストラム平均を求める。フレームtにおける次数iのケプストラムを$c(i,t)$と表すとき、フレームtにおける音声フレームのケプストラム平均$\tilde{C}_{speech}(i,t)$は、例えば(19)式により求めることができる。

$$\tilde{C}_{speech}(i,t) = \begin{cases} \eta \cdot \tilde{C}_{speech}(i,t-1) + (1-\eta) \cdot c(i,t) & \text{if frame t is speech} \\ C_{speech}(i,t-1) & \text{otherwise} \end{cases} \quad (19)$$

ここで、ηは音声フレームのケプストラム平均を求める際の平滑化係数であり、１．０よりも小さく１．０に近い値を設定すればよい。また、フレームtにおける非音声フレームのケプストラム平均$\tilde{C}_{nonspeech}(i,t)$は、例えば(20)式により求めることができる。ここで、ηは非音声フレームのケプスト

ラム平均を求める際の平滑化係数であり、１．０よりも小さく１．０に近い値を設定すればよい。

$$\tilde{C}_{nonspeech}(i,t) = \begin{cases} \eta \cdot \tilde{C}_{nonspeech}(i,t-1) + (1-\eta) \cdot c(i,t) & \text{if frame t is not speech} \\ \tilde{C}_{nonspeech}(i,t-1) & \text{otherwise} \end{cases}$$

(20)

音声フレームのケプストラム平均は、乗法性歪み$H^*(\omega)$のケプストラム表現であり、話者に依存する。一方、非音声フレームのケプストラム平均は、マイク特性などの入力系の乗法性歪み$H_{Mic}(\omega)$に依存する。

正規化ステップ：観測ケプストラム$c(i,t)$から、(21)式に従って、音声フレームにおいては音声フレームのケプストラム平均$\tilde{C}_{speech}(i,t)$を、非音声フレームにおいては非音声フレームのケプストラム平均$\tilde{C}_{nonspeech}(i,t)$を話者毎に引き、正規化ケプストラム$\hat{C}(i,t)$を求めることにより、観測スペクトルを正規化する。

$$\hat{C}(i,t) = \begin{cases} c(i,t) - \tilde{C}_{speech}(i,t) & \text{if frame t is speech} \\ c(i,t) - \tilde{C}_{nonspeech}(i,t) & \text{otherwise} \end{cases} \quad (21)$$

尚、Ｅ－ＣＭＮ法に関する発明について本願出願人は、既に日本国出願を済ませている（特願平０９－０５１５７８号、１９９７年 3月１１日）。

Ｅ－ＣＭＮ法は、様々な乗法性歪みの積を１０単語程度の少量の音声から、音声区間のケプストラム平均として推定し、それを入力ケプストラムから引くという方法である。不特定話者音素モデルを観測されたスペクト

ルから求められたケプストラムではなく、Ｅ－ＣＭＮ法により正規化されたケプストラムを用いて作成しておくことにより、様々な乗法性歪みを一括して補正することが可能であることが明らかにされている。

最後に、音源既知および音源未知の加法性雑音、乗法性歪みが存在する実環境におけるロバストな音声認識手法として、ＮＬＭＳ－ＶＡＤ法、ＣＳＳ法、Ｅ－ＣＭＮ法を組み合わせる手法について述べる。図12に本組み合わせ手法に従って構成した演算回路のブロック図を示す。まず、第１の回路１０１においてＮＬＭＳ－ＶＡＤ法により、入力音声から音源既知の加法性雑音$\bar{E}(\omega;t)$が除去された波形信号が生成される。次に、第２の回路１０２においてこの波形信号に対して、フーリエ変換が施された後、ＣＳＳ法により音源未知の加法性雑音$\bar{N}(\omega;t)$が除去されたスペクトルの時系列が生成される。

さらに、第３の回路１０３においてこのスペクトルの時系列はケプストラムの時系列に変換され、Ｅ－ＣＭＮ法により正規化されたケプストラムの時系列に変換される。最後に、第４の回路１０４においてケプストラムの時系列は、公知のビタビアルゴリズムにより、不特定話者用に予め作成された音素モデルと照合され、音声認識結果が出力される。

この際に使用される不特定話者用音素モデルは、先述のようにＥ－ＣＭＮ法により正規化されたケプストラムを用いて作成されている必要がある。尚、Ｅ－ＣＭＮ法で必要な音声フレーム・非音声フレームの区別は、ＮＬＭＳ－ＶＡＤ法に組み込まれたＶＡＤの結果をそのまま用いればよい。

次に、本組み合わせ手法の効果についてまとめる。単一指向性マイクを２０００ｃｃの自動車の運転席サンバイザーに設置し、男性２名女性２名が各々好みの位置にセットした運転席に座って発声した５２０単語（ＡＴ

Ｒ音声データベースＣセット）の音声（データ１）を収録した。音声区間の前後に２５０ｍｓずつの無音区間が付属するように手動で切り出しを行った。また、アイドリング、時速６０ｋｍ、時速１００ｋｍの走行状態で、５種類の音楽ソース（ポップス、ジャズ、ロック、クラシック、落語）を順番にカーオーディオで再生し、音楽ソースの左右チャンネルを混合した信号（データ２）と、マイク入力信号（データ３）のペアを同時に録音した。カーオーディオの出力ボリュームは、各走行状態毎に男性１名が快適と感じる音量にセットした。

アイドリング、時速６０ｋｍ、時速１００ｋｍでの音響エコーのマイクへの最大入力レベルはそれぞれ、６０．７ｄＢＡ、６５．９ｄＢＡ、７０．６ｄＢＡであった。データ１とデータ３を計算機上で加算して評価データを作成した。データ２は、ＮＬＭＳ－ＶＡＤ法の遠端入力として使用した。認識には、環境独立な５４音素の不特定話者用Tied-MixtureＨＭＭモデル（４０名の音声データから作成）を用いた。分析条件は８ｋＨｚサンプリング、フレーム長３２ｍｓ、フレームシフト１０ｍｓで、特徴パラメータは、１０次ＭＦＣＣ、１０次ΔＭＦＣＣ、Δエネルギーであり、ＨＭＭモデルが共有する正規分布の数は、それぞれ２５６、２５６、６４である。

不特定話者、５２０単語の認識タスクで、アイドリング、時速６０ｋｍ、時速１００ｋｍの走行状態で、スピーカー出力音が存在しない場合（w/o Speaker Out）、スピーカー出力音が存在するが、ＮＬＭＳ－ＶＡＤ法を行わない場合（w/ Speaker Out w/o NLMS-VAD）、スピーカー出力音が存在し、ＮＬＭＳ－ＶＡＤ法を行う場合（w/ Speaker Out w/ NLMS-VAD）の認識性能（５種類の音楽ソースの平均）およびＲＲＥを図１３に示す。

いずれの走行状態でも８０％以上のＲＲＥが得られた。また、ＮＬＭＳ－ＶＡＤ法でも回復できない誤認識率は、アイドリング、時速６０ｋｍ、

時速１００ｋｍでそれぞれ０．７％、２．１％、１．８％と僅かであり、先の組み合わせ法の有効性が確認できた。

（第１実施形態）

第１実施形態の音声処理装置の回路構成を図１４に示す。以下に述べる個々の手段は周知の回路、たとえば、デジタル回路、コンピュータやデジタルプロセッサの演算処理により実現する回路を使用できるので、当業者であれば、図１３により音声処理装置を製作できるであろう。まず、サンプル毎の処理について説明する。音源信号１１はスピーカー出力信号１３として、スピーカーから出力される。供給手段ａａ７は、現時点のインパルス応答（ＦＩＲフィルターの係数）を保持し、生成手段ａａ４に現時点のインパルス応答１６を供給する。

音源信号１１は生成手段ａａ４に送られ、生成手段ａａ４でＦＩＲフィルターにより、疑似音響エコー信号１５が生成される。除去手段ａａ５において、疑似音響エコー信号１５はマイク入力信号１２から減じられ、音源エコーキャンセル信号１４が生成される。更新手段ａａ３において、音源信号１１と音源エコーキャンセル信号１４と供給手段ａａ７が保持する現時点のインパルス応答１６から、インパルス応答を更新し、更新インパルス応答１７を生成して、それを供給手段ａａ７に供給する。

供給手段ａａ７は後述の音声判定フラグ１８がＯＦＦの間にのみ、更新手段ａａ３から供給された更新インパルス応答１７を新しい現時点のインパルス応答として保持し、音声判定フラグ１８がＯＮの間は、更新手段ａａ３から供給された更新インパルス応答１７を棄却する。除去手段ａａ５で生成された音響エコーキャンセル信号１４は、判定手段ａａ１にも送られる。

次に、フレーム毎の処理について説明する。サンプル毎に判定手段ａａ

１に送られた音響エコーキャンセル信号１４は判定手段ａａ１においてバッファに格納され、１フレーム分たまった段階で判定手段ａａ１でマイク入力手段（不図示）に音声が存在するかどうかの判定が行われ、音声判定フラグ１８が出力される。肯定判定の場合（音声の存在が検出される場合）は、音声判定フラグの値はＯＮであるという。否定判定の場合（音声の存在が検出されない場合）は、音声判定フラグの値はＯＦＦであるという。

この判定処理では、音響エコーキャンセル信号１４の時間情報および周波数情報を利用して、音声信号がマイク入力信号１２に含まれているかどうかを判定する。判定の方法としては、音源未知の加法性雑音に重畳した音声を検出できるアルゴリズムVAD(Voice Activity Detection)を用いると良い。例えば、Recommendation GSM 06.32では、信号のＬＰＣ分析後の残差エネルギーをしきい値と比較することにより音声の検出を行うが、音源未知の加法性雑音のエネルギーレベルに合わせてしきい値を変動させることができるため、音源未知の加法性雑音と音声を分離することができる。

しきい値の適応化を行う際には、スペクトルの定常性（周波数情報）とピッチ性（時間情報）を利用している。上記判定処理の結果、音声判定フラグ１８がＯＦＦの場合は、制御手段ａａ６が供給手段ａａ７に保持された現時点インパルス応答１６を取り出し、所望のインパルス応答として保存手段ａａ２に格納する。

一方、音声判定フラグ１８がＯＮの場合は、供給手段ａａ７が保持しているインパルス応答が所望の値から乖離している可能性があるため、制御手段ａａ６は保存手段ａａ２から保存インパルス応答を１つ取り出し、供給手段ａａ７が保持しているインパルス応答に上書きする。保存手段ａａ２は、１つ以上のインパルス応答を保存できるＦＩＦＯであればよい。

（第２実施形態）

図１５に第２実施形態の基本構成を示す。まず、第１の実施形態で説明した図１４の構成を持つ音声処理装置１００は、音源信号１１およびマイク入力信号１２を用いて、マイク入力信号１２に含まれる音響エコーをキャンセルし、音響エコーキャンセル信号１４を生成する。次に、スペクトル計算手段ｂｂ１において、一定フレーム周期毎にフーリエ変換により音響エコーキャンセル信号１４のスペクトル２１を算出する。

スペクトル２１はケプストラム計算手段ｂｂ４に送られ、一定フレーム毎のケプストラム２４に変換される。照合手段ｂｂ５では、一定フレーム毎のケプストラム２４を用いて、照合が行われ、認識結果２５が出力される。照合に当たっては、周知の隠れマルコフモデルによる手法または周知の動的計画法による手法または周知のニューラルネットワークによる手法のいずれかを用いればよい。

（第３実施形態）

図１６に第３実施形態の基本構成ををを示す。図１５の第２実施形態と同様の個所には同一の符号を付している。まず、図１４の構成を持つ音声処理装置１００は、音源信号１１およびマイク入力信号１２を用いて、マイク入力信号１２に含まれる音響エコーをキャンセルし、音響エコーキャンセル信号１４を生成する。次に、スペクトル計算手段ｂｂ１において、一定フレーム周期毎にフーリエ変換により音響エコーキャンセル信号１４のスペクトル２１を算出する。スペクトル２１は、スペクトル平均計算手段ｂｂ２に送られ、一定フレーム毎に(8)式によりスペクトル平均２２が求められる。

一方、スペクトル計算手段ｂｂ１で求められたスペクトル２１は、スペクトル平均減算手段ｂｂ３に供給され、(9)式によりスペクトル平均２２が

減じられ、雑音除去スペクトル２３が求められる。雑音除去スペクトル２３はケプストラム計算手段ｂｂ４に送られ、一定フレーム毎のケプストラム２４に変換される。照合手段ｂｂ５では、一定フレーム毎のケプストラム２４を用いて、照合が行われ、認識結果２５が出力される。照合に当たっては、周知の隠れマルコフモデルによる手法または周知の動的計画法による手法または周知のニューラルネットワークによる手法のいずれかを用いればよい。

（第４実施形態）

図１７に第４実施形態の基本構成を示す。図１７において第２または第３実施形態と同様の個所には同一の符号を付している。まず、図１４の構成を持つ音声処理装置１００は、音源信号１１およびマイク入力信号１２を用いて、マイク入力信号１２に含まれる音響エコーをキャンセルし、音響エコーキャンセル信号１４を生成する。次に、スペクトル計算手段ｂｂ１において、一定フレーム周期毎にフーリエ変換により音響エコーキャンセル信号１４のスペクトル２１を算出する。スペクトル２１は、スペクトル平均計算手段ｂｂ２に送られ、一定フレーム毎に(8)式によりスペクトル平均２２が求められる。

一方、スペクトル計算手段ｂｂ１で求められたスペクトル２１は、スペクトル平均減算手段ｂｂ３に供給され、(9)式によりスペクトル平均２２が減じられ、雑音除去スペクトル２３が求められる。雑音除去スペクトル２３はケプストラム計算手段ｂｂ４に送られ、一定フレーム毎のケプストラム２４に変換される。

ケプストラム２４はケプストラム平均計算手段ｃｃ１に送られ、そこでケプストラム平均３１が求められる。ケプストラム平均の算出に当たっては、例えば(19)式、(20)式を用いればよい。次に、ケプストラム平均減算手

段ｃｃ２において、ケプストラム２４からケプストラム平均３１が減算し、正規化ケプストラム３２が算出される。尚、減算に当たっては、(21)式を用いればよい。照合手段ｂｂ５では、一定フレーム毎の正規化ケプストラム３２を用いて、照合が行われ、認識結果２５が出力される。照合に当たっては、周知の隠れマルコフモデルによる手法または周知の動的計画法による手法または周知のニューラルネットワークによる手法のいずれかを用いればよい。

（第５実施形態）

図１８に第５実施形態の基本構成を示す。図１８において第２、第３または第４実施形態と同様の個所には同一の符号を付している。まず、図１４の構成を持つ音声処理装置１００は、音源信号１１およびマイク入力信号１２を用いて、マイク入力信号１２に含まれる音響エコーをキャンセルし、音響エコーキャンセル信号１４を生成する。次に、スペクトル計算手段ｂｂ１において、一定フレーム周期毎にフーリエ変換により音響エコーキャンセル信号１４のスペクトル２１を算出する。スペクトル２１は、ケプストラム計算手段ｂｂ４に送られ、一定フレーム毎のケプストラム２４に変換される。ケプストラム２４はケプストラム平均計算手段ｃｃ１に送られ、そこでケプストラム平均３１が求められる。ケプストラム平均の算出に当たっては、例えば(19)式、(20)式を用いればよい。

次に、ケプストラム平均減算手段ｃｃ２において、ケプストラム２４からケプストラム平均３１を減算し、正規化ケプストラム３２が算出される。尚、減算に当たっては、(21)式を用いればよい。照合手段ｂｂ５では、一定フレーム毎の正規化ケプストラム３２を用いて、照合が行われ、認識結果２５が出力される。照合に当たっては、周知の隠れマルコフモデルによる手法または周知の動的計画法による手法または衆知のニューラルネット

ワークによる手法のいずれかを用いればよい。

（第６実施形態）

図１９に第６実施形態の基本構成を示す。図１６においても第２、第３第４または第５実施形態と同様の個所には同一の符号を付している。まず、図１４の構成を持つ音声処理装置１００により、音源信号１１およびマイク入力信号１２を用いて、マイク入力信号１２に含まれる音響エコーをキャンセルし、音響エコーキャンセル信号１４を生成する。次に、スペクトル計算手段ｂｂ１において、一定フレーム周期毎にフーリエ変換により音響エコーキャンセル信号１４のスペクトル２１を算出する。スペクトル２１は、ケプストラム計算手段ｂｂ４に送られ、一定フレーム毎のケプストラム２４に変換される。

ケプストラム２４はケプストラム平均計算手段ｃｃ１に送られ、そこでケプストラム平均３１が求められる。ケプストラム平均の算出に当たっては、例えば(19)式、(20)式を用いればよい。次に、ケプストラム平均減算手段ｃｃ２において、ケプストラム２４からケプストラム平均３１を減算し、正規化ケプストラム３２が算出される。尚、減算に当たっては、(21)式を用いればよい。照合手段ｂｂ５では、一定フレーム毎の正規化ケプストラム３２を用いて、照合が行われ、認識結果２５が出力される。照合に当たっては、周知の隠れマルコフモデルによる手法または周知の動的計画法による手法または公知のニューラルネットワークによる手法のいずれかを用いればよい。

請求の範囲

1．音響エコーの伝達経路を模擬する現時点のインパルス応答および音源信号に基づき疑似音響エコー信号を生成する生成手段と、

現時点のインパルス応答を保持し、前記生成手段に供給する供給手段と、

マイク入力信号から前記疑似音響エコー信号を減算することにより音響エコー成分を除去し、音響エコーキャンセル信号を生成する除去手段と、

前記音源信号と前記音響エコーキャンセル信号と前記供給手段が保持している現時点のインパルス応答を用いて継続的にインパルス応答を更新し、更新されたインパルス応答を前記供給手段に供給する更新手段と、

前記音響エコーキャンセル信号の時間情報および周波数情報を利用して、マイク入力信号に音声が含まれているか否かをフレーム毎に判定する判定手段と、

1つ以上のインパルス応答を保存する保存手段と、

前記判定手段の判定結果が否定判定のフレームでは前記供給手段が保持している現時点のインパルス応答を前記保存手段に保存し、肯定判定のフレームでは前記保存手段に保存されているインパルス応答の1つを取り出して、前記供給手段に供給する制御手段と

を具えたことを特徴とする音声処理装置。

2．請求の範囲第1項に記載の音声処理装置において、前記音響エコーがキャンセルされた後の信号を音声認識に用いることを特徴とする音声処理装置。

3．請求の範囲第2項に記載の音声処理装置において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペ

クトルを求める手段と、当該得られたスペクトルに基づき各フレーム毎に連続的にスペクトル平均を求める手段と、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信号から各フレーム毎に計算されたスペクトルから連続的に減算することにより、音源未知の加法性雑音を除去する手段とをさらに具えたことを特徴とする音声処理装置。

4．請求の範囲第2項に記載の音声処理装置において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルから各フレーム毎に連続的にスペクトル平均を求める手段と、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信号から各フレーム毎に計算されたスペクトルを連続的に減算する手段と、当該減算する手段により、音源未知の加法性雑音が除去されたスペクトルからケプストラムを求める手段と、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求める手段と、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算して、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正する手段をさらに具えたことを特徴とする音声処理装置。

5．請求の範囲第2項に記載の音声処理装置において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求める手段と、当該得られたスペクトルからケプストラムを求める手段と、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求める手段と、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケ

プストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算することにより、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正する手段とをさらに具えたことを特徴とする。

6．フーリエ変換により各フレーム毎にスペクトルを求める手段と、

当該得られたスペクトルからケプストラムを求める手段と、

当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求める手段と、

話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算することによりマイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正する手段と

を具えたことを特徴とする音声処理装置。

7．音響エコーの伝達経路を模擬する現時点のインパルス応答および音源信号に基づき疑似音響エコー信号を生成する生成ステップと、

現時点のインパルス応答を保持し、前記生成手段に供給する供給ステップと、

マイク入力信号から前記疑似音響エコー信号を減算することにより音響エコー成分を除去し、音響エコーキャンセル信号を生成する除去ステップと、

前記音源信号と前記音響エコーキャンセル信号と前記供給ステップで保持している現時点のインパルス応答を用いて継続的にインパルス応答を更新し、更新されたインパルス応答を前記供給ステップに供給する更新ステップと、

前記音響エコーキャンセル信号の時間情報および周波数情報を利用して、マイク入力信号に音声が含まれているか否かをフレーム毎に判定する判定ステップと、

１つ以上のインパルス応答を保存する保存ステップと、

前記判定ステップの判定結果が否定判定のフレームでは前記供給ステップが保持している現時点のインパルス応答を前記保存ステップで保存し、肯定判定のフレームでは前記保存ステップで保存されているインパルス応答の１つを取り出して、前記供給ステップに供給する制御ステップと

を具えたことを特徴とする音声処理方法。

８．請求の範囲第７項に記載の音声処理方法において、前記音響エコーがキャンセルされた後の信号を音声認識に用いることを特徴とする音声処理方法。

９．請求の範囲第８項に記載の音声処理方法において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求めるステップと、当該得られたスペクトルに基づき各フレーム毎に連続的にスペクトル平均を求めるステップと、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信号から各フレーム毎に計算されたスペクトルを連続的に減算することにより、音源未知の加法性雑音を除去するステップとをさらに具えたことを特徴とする音声処理方法。

１０．請求の範囲第８項に記載の音声処理方法において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求めるステップと、当該得られたスペクトルから各フレーム毎に連続的にスペクトル平均を求めるステップと、当該得られたスペクトル平均を前記音響エコーがキャンセルされた後の信号から各フレーム毎に計算されたスペクトルから連続的に減算することにより、音源未知の加法性

雑音を除去するステップと、当該加法性雑音が除去されたスペクトルからケプストラムを求めるステップと、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求めるステップと、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算して、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正するステップをさらに具えたことを特徴とする音声処理方法。

１１．請求の範囲第８項に記載の音声処理方法において、前記音響エコーがキャンセルされた後の信号から、フーリエ変換により各フレーム毎にスペクトルを求めるステップと、当該得られたスペクトルからケプストラムを求めるステップと、当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求めるステップと、話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、その話者の非音声フレームのケプストラムからは非音声フレームのケプストラム平均を減算することにより、マイク特性や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正するステップとをさらに具えたことを特徴とする音声処理方法。

１２．フーリエ変換により各フレーム毎にスペクトルを求めるステップと、

当該得られたスペクトルからからケプストラムを求めるステップと、

当該得られたケプストラムの音声フレームのケプストラム平均および非音声フレームのケプストラム平均を話者毎に別々に求めるステップと、

話者毎に音声フレームのケプストラムからはその話者の音声フレームのケプストラム平均を減算し、非音声フレームのケプストラムからはその話者の非音声フレームのケプストラム平均を減算することによりマイク特性

や口からマイクまでの空間伝達特性に依存する乗法性歪みを補正するステップと

を具えたことを特徴とする音声処理方法。

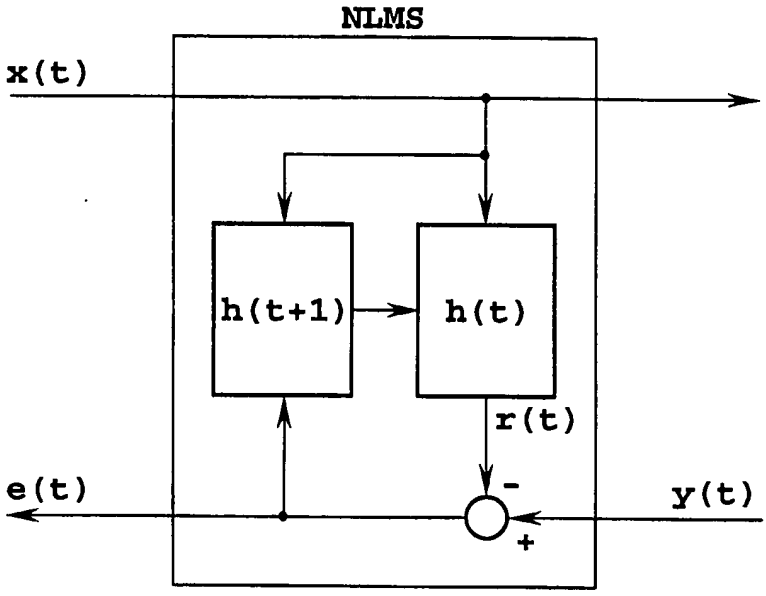


FIG.1

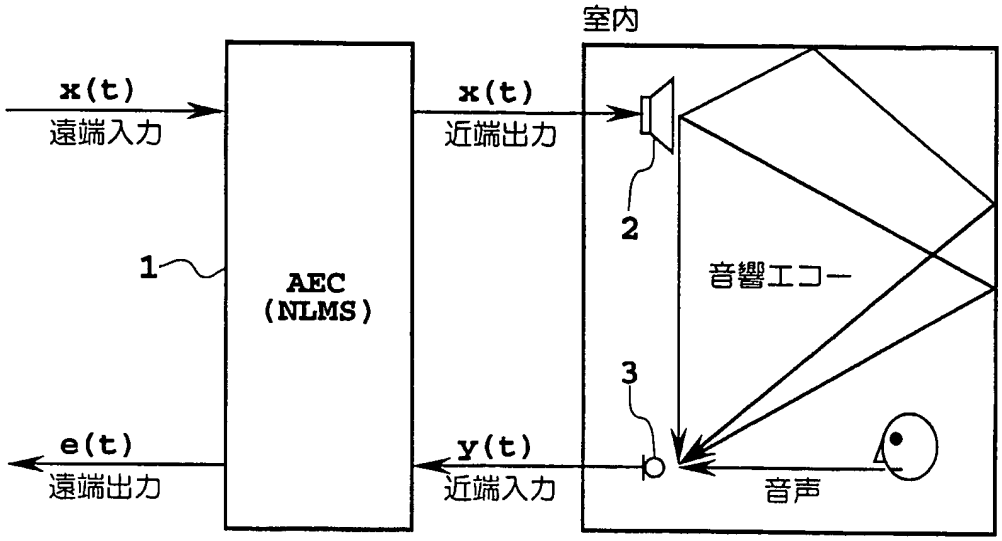

室内
x(t)
遠端入力
x(t)
近端出力
1
AEC
(NLMS)
2
音響エコー
3
e(t)
遠端出力
y(t)
近端入力
音声

FIG.2

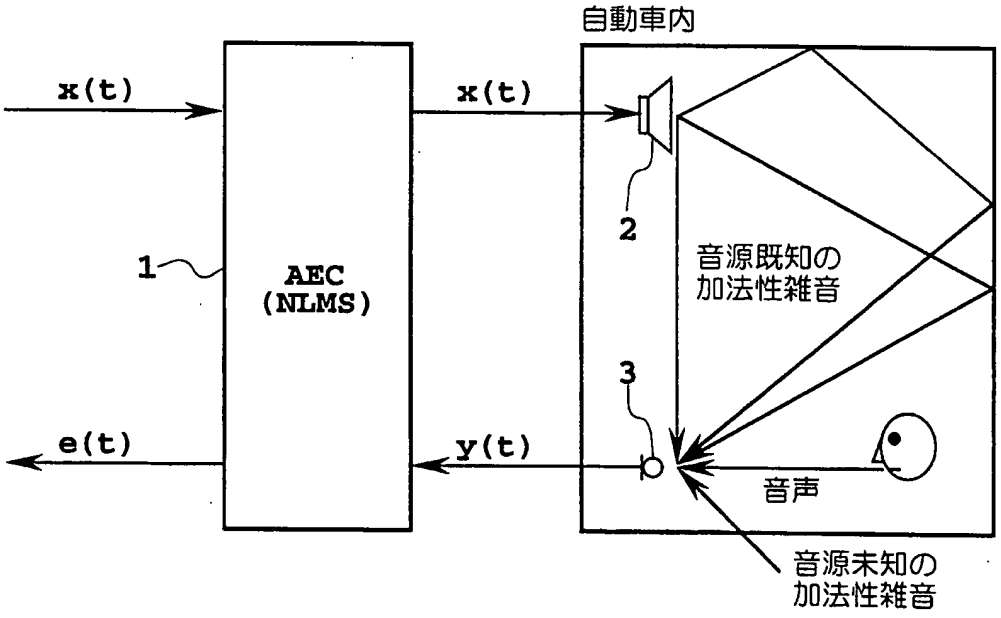

自動車内
x(t)
x(t)
1
AEC
(NLMS)
2
音源既知の
加法性雑音
3
e(t)
y(t)
音声
音源未知の
加法性雑音

FIG.3

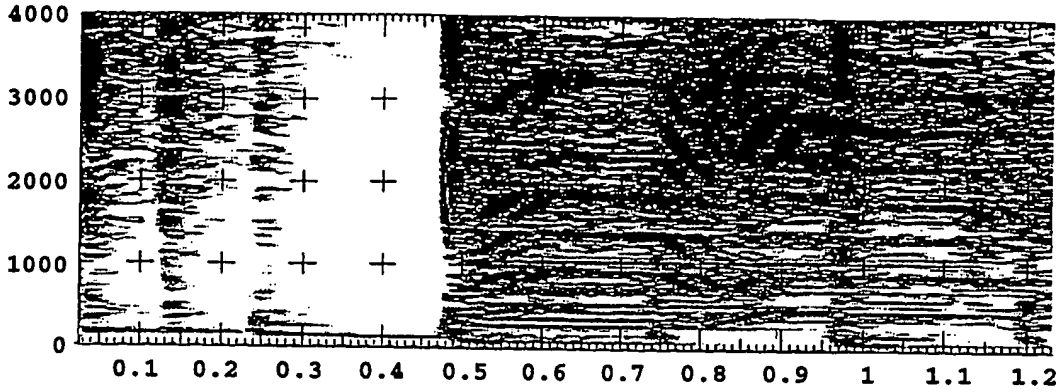

4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2

FIG.4A

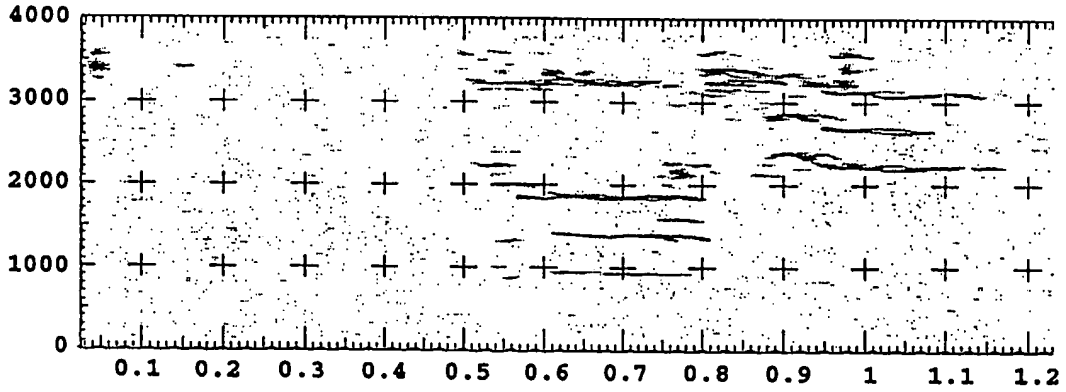

4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2

FIG.4B

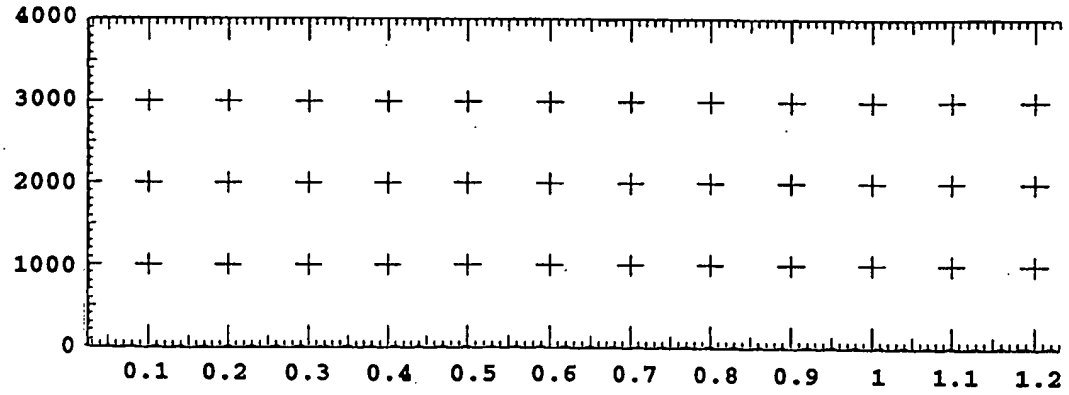


FIG.4C

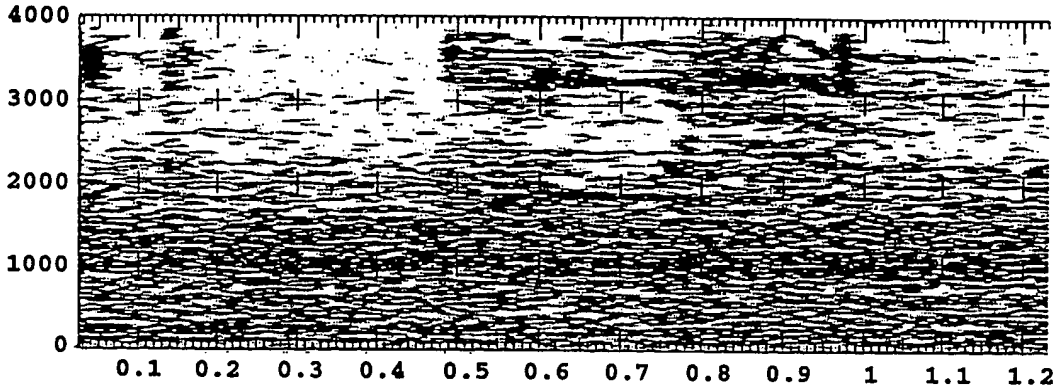


FIG.4D

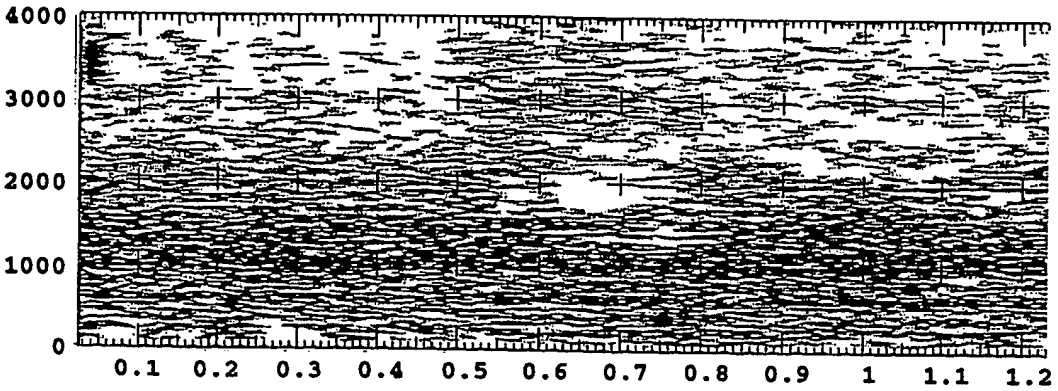


FIG.4E

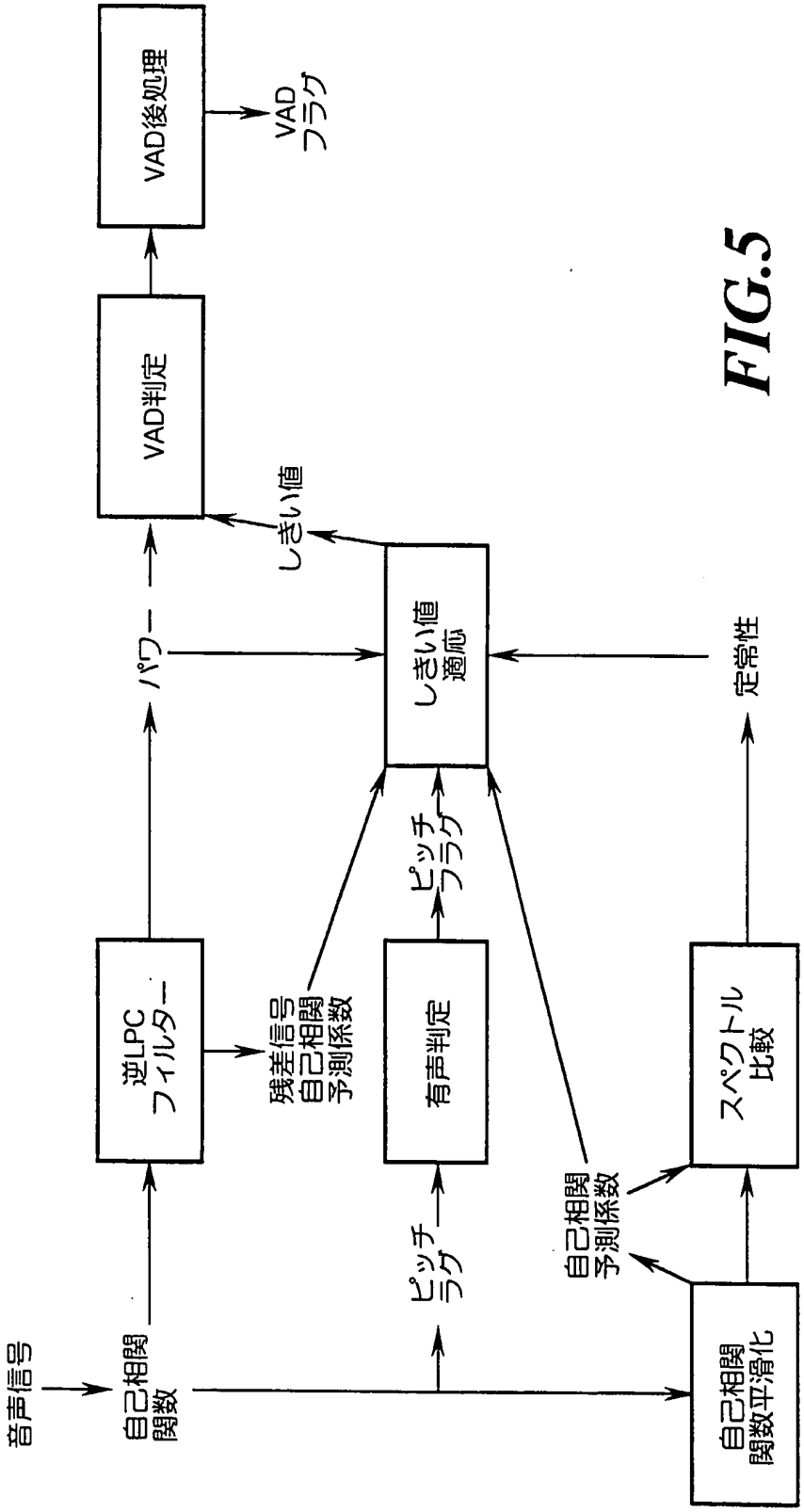

音声信号
自己相関関数
逆LPCフィルター
パワー
残差信号自己相関予測係数
ピッチラグ
有声判定
ピッチラグ
自己相関予測係数
自己相関関数平滑化
スペクトル比較
定常性
しきい値適応
しきい値
VAD判定
VAD後処理
VADフラグ

FIG.5

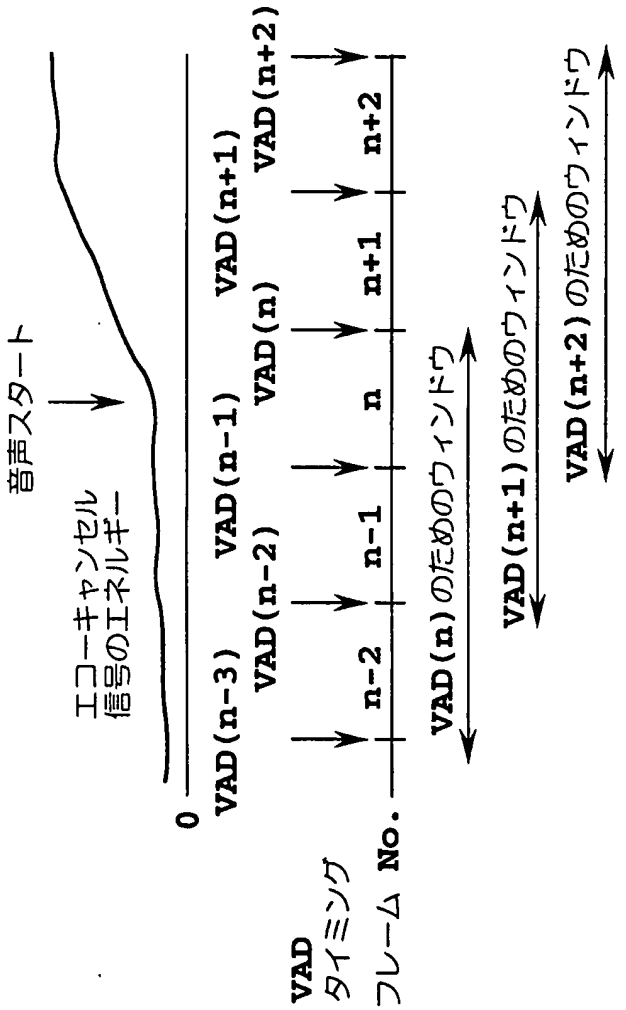

音声スタート
エコーキャンセル
信号のエネルギー
0
VAD
タイミング
フレーム No.
VAD(n-3)
VAD(n-2)
VAD(n-1)
VAD(n)
VAD(n+1)
VAD(n+2)
n-2
n-1
n
n+1
n+2
VAD(n)のためのウィンドウ
VAD(n+1)のためのウィンドウ
VAD(n+2)のためのウィンドウ

FIG.6

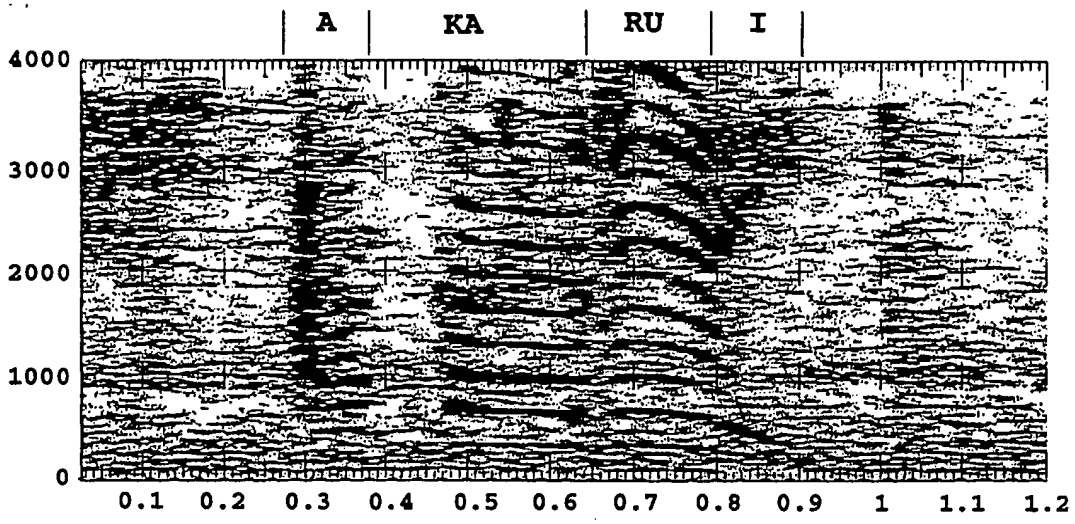


FIG.7A

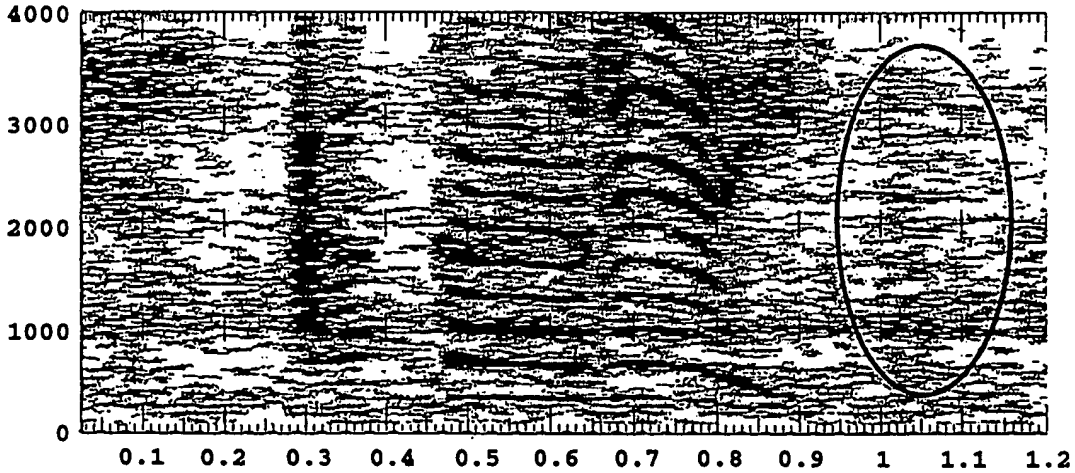


FIG.7B

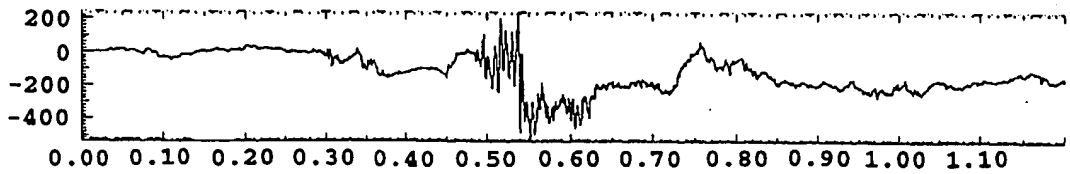


FIG.7C

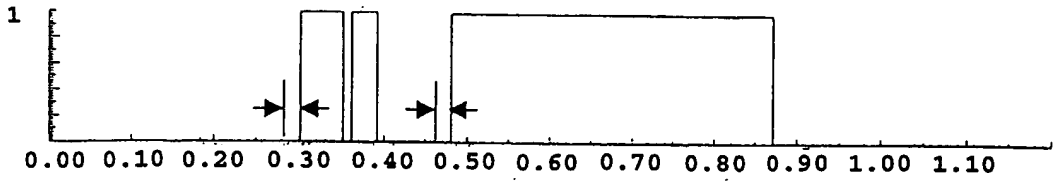


FIG.7D

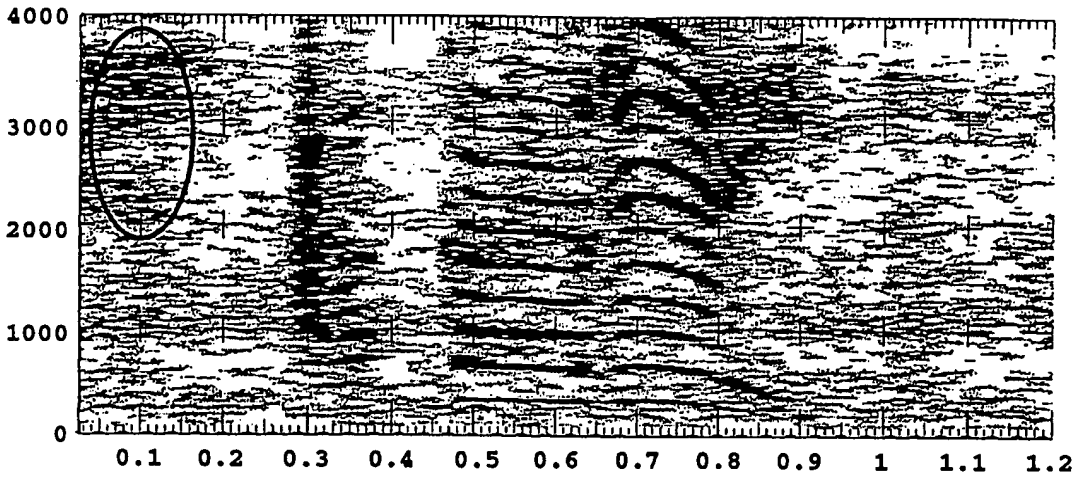


FIG.7E

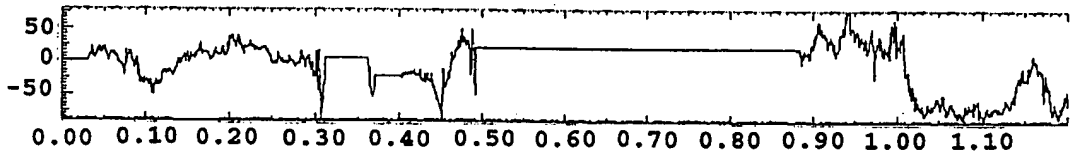


FIG.7F

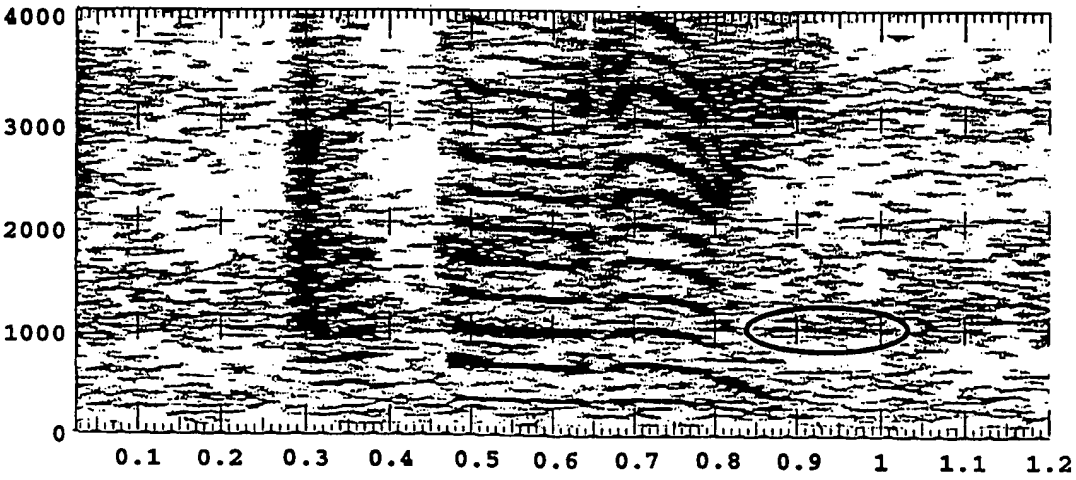


FIG.7G

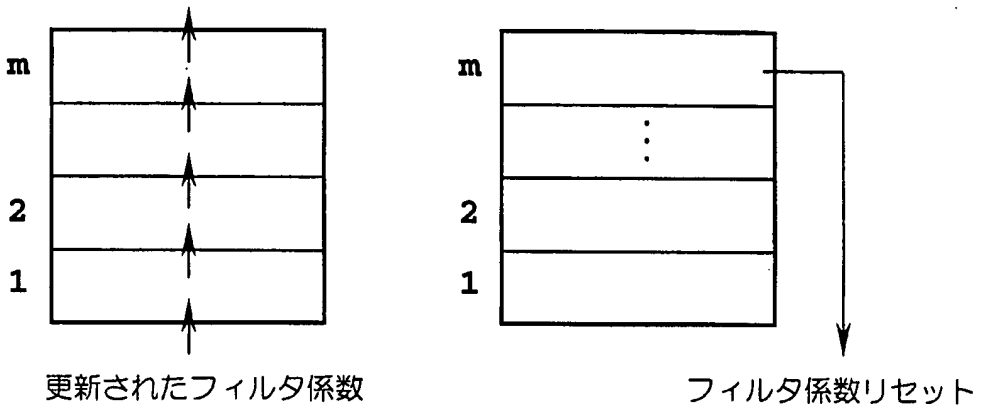


FIG.8

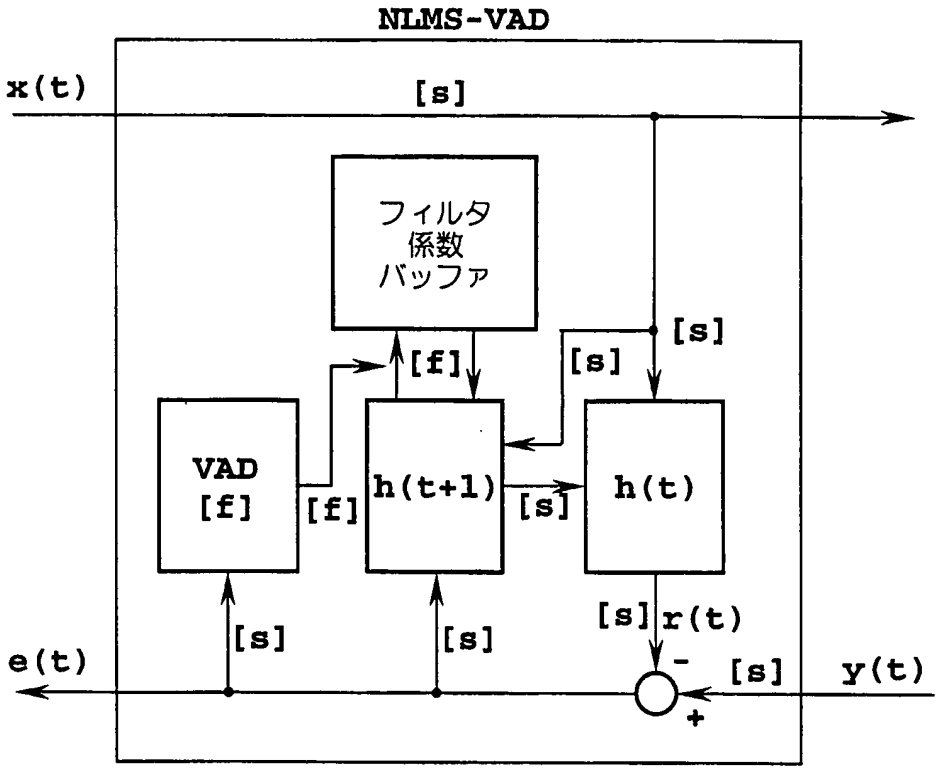


FIG.9

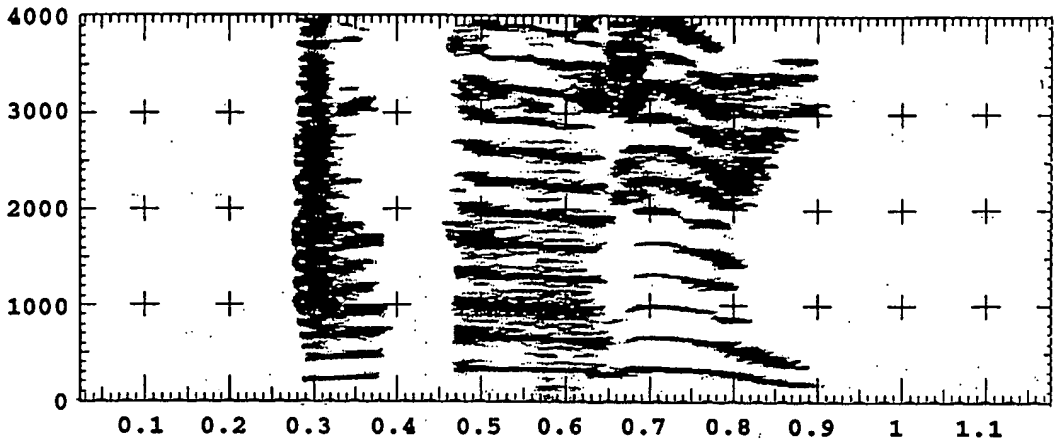


FIG.10A

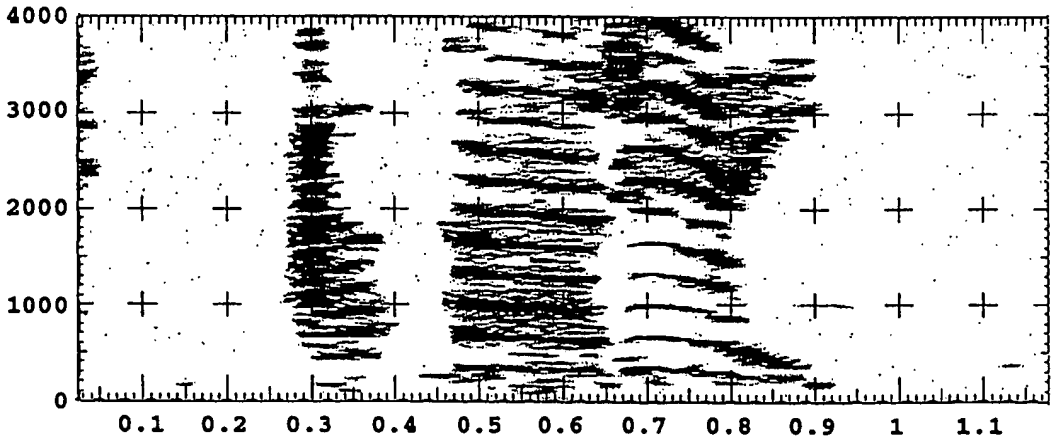


FIG.10B

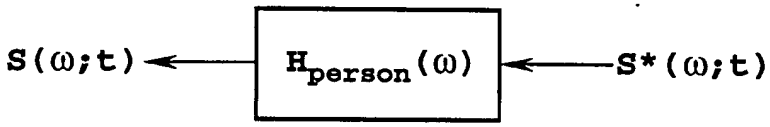


FIG.11

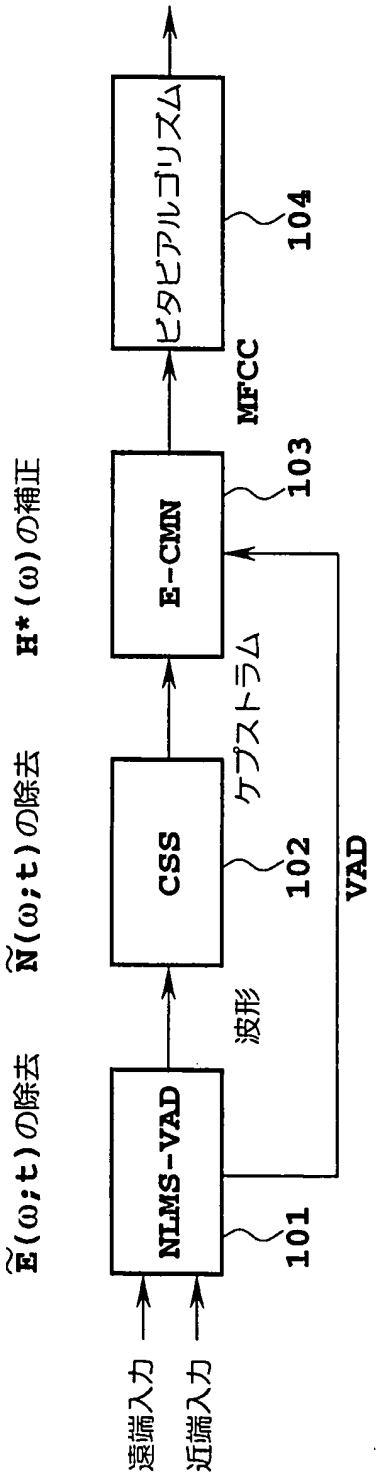

遠端入力
近端入力
NLMS-VAD
101
$\tilde{E}(\omega;t)$の除去
波形
CSS
102
$\tilde{N}(\omega;t)$の除去
ケプストラム
E-CMN
103
$H^*(\omega)$の補正
MFCC
ビタビアルゴリズム
104
VAD

FIG.12

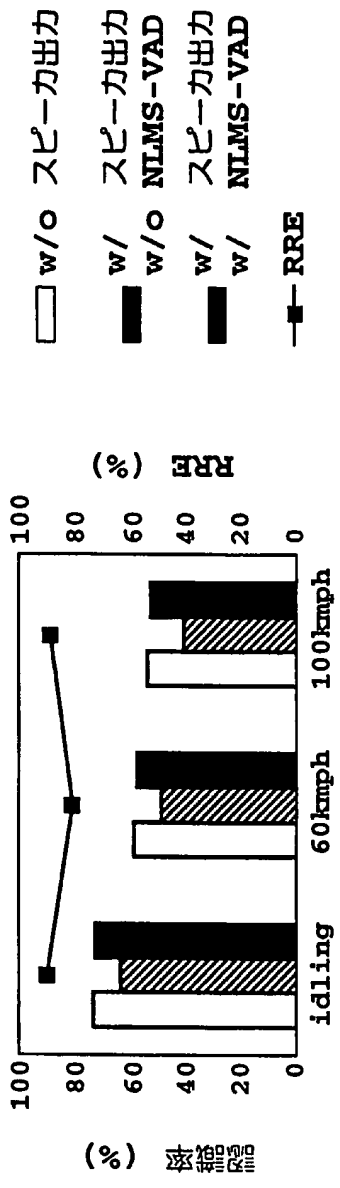

w/o スピーカ出力
w/ スピーカ出力 w/o NLMS-VAD
w/ スピーカ出力 w/ NLMS-VAD
RRE
RRE (%)
100
80
60
40
20
0
idling
60kmph
100kmph
騒音レベル (%)

FIG.13

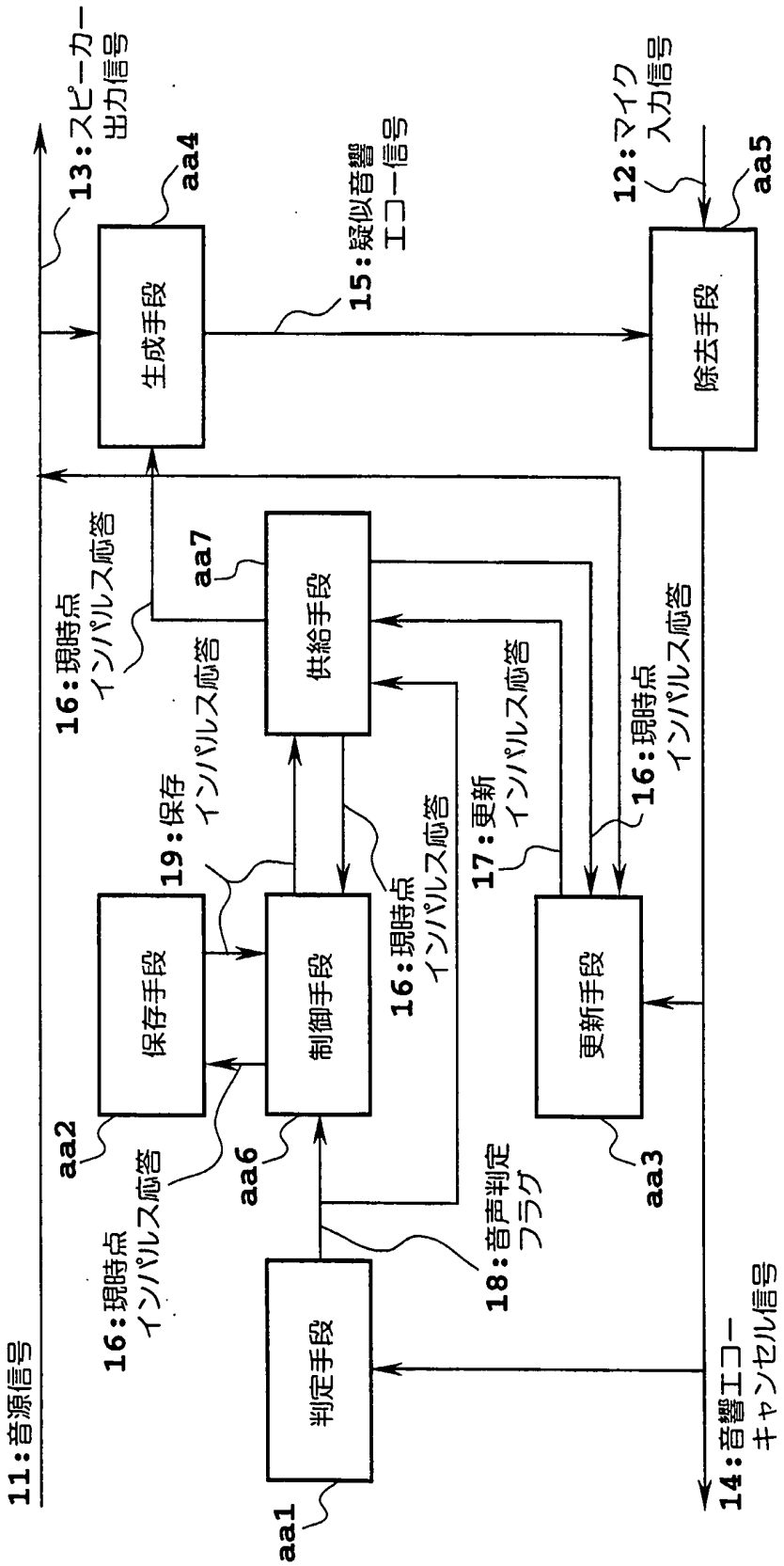

11:音源信号
13:スピーカー出力信号
12:マイク入力信号
14:音響エコーキャンセル信号
15:疑似音響エコー信号
16:現時点インパルス応答
17:更新インパルス応答
18:音声判定フラグ
19:保存インパルス応答
生成手段
除去手段
供給手段
保存手段
制御手段
更新手段
判定手段
aa1
aa2
aa3
aa4
aa5
aa6
aa7

FIG.14

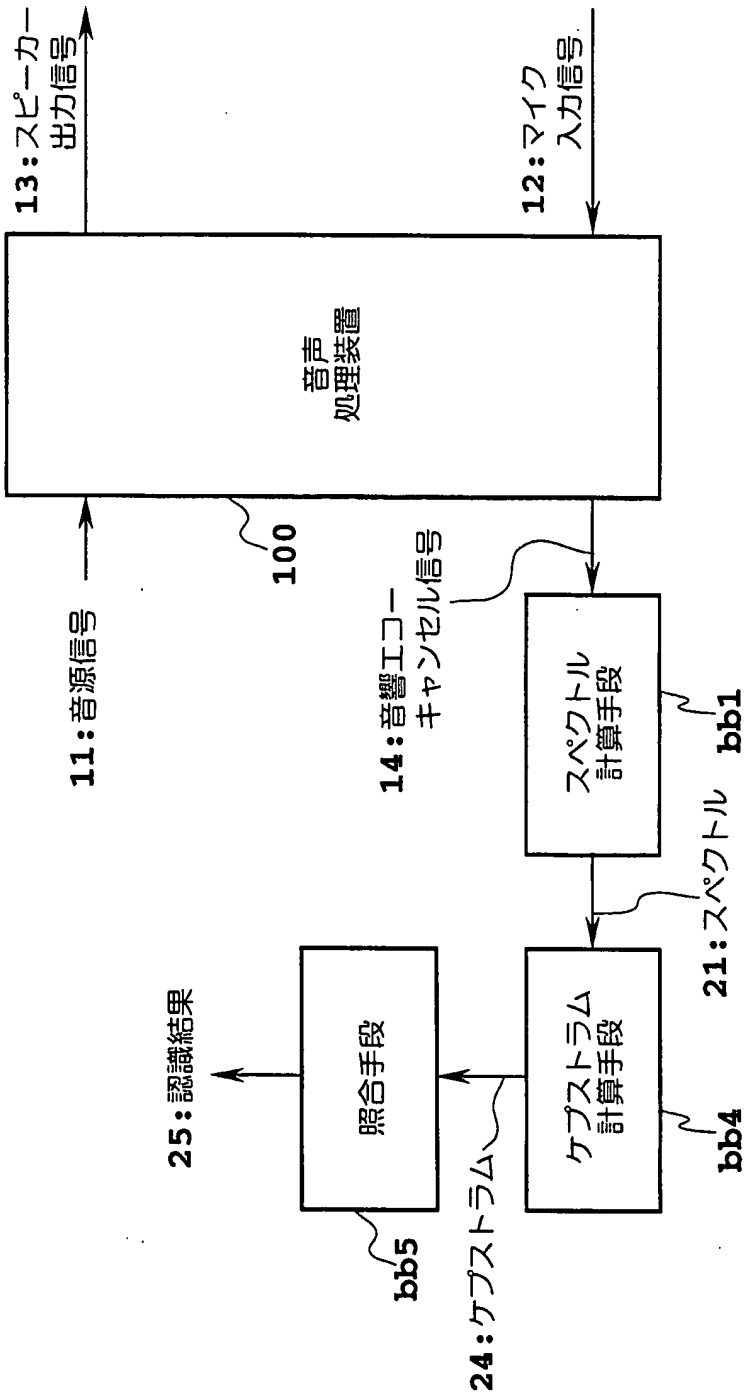

13:スピーカー出力信号
12:マイク入力信号
音声処理装置
100
11:音源信号
14:音響エコーキャンセル信号
スペクトル計算手段
bb1
21:スペクトル
ケプストラム計算手段
bb4
24:ケプストラム
照合手段
bb5
25:認識結果

FIG.15

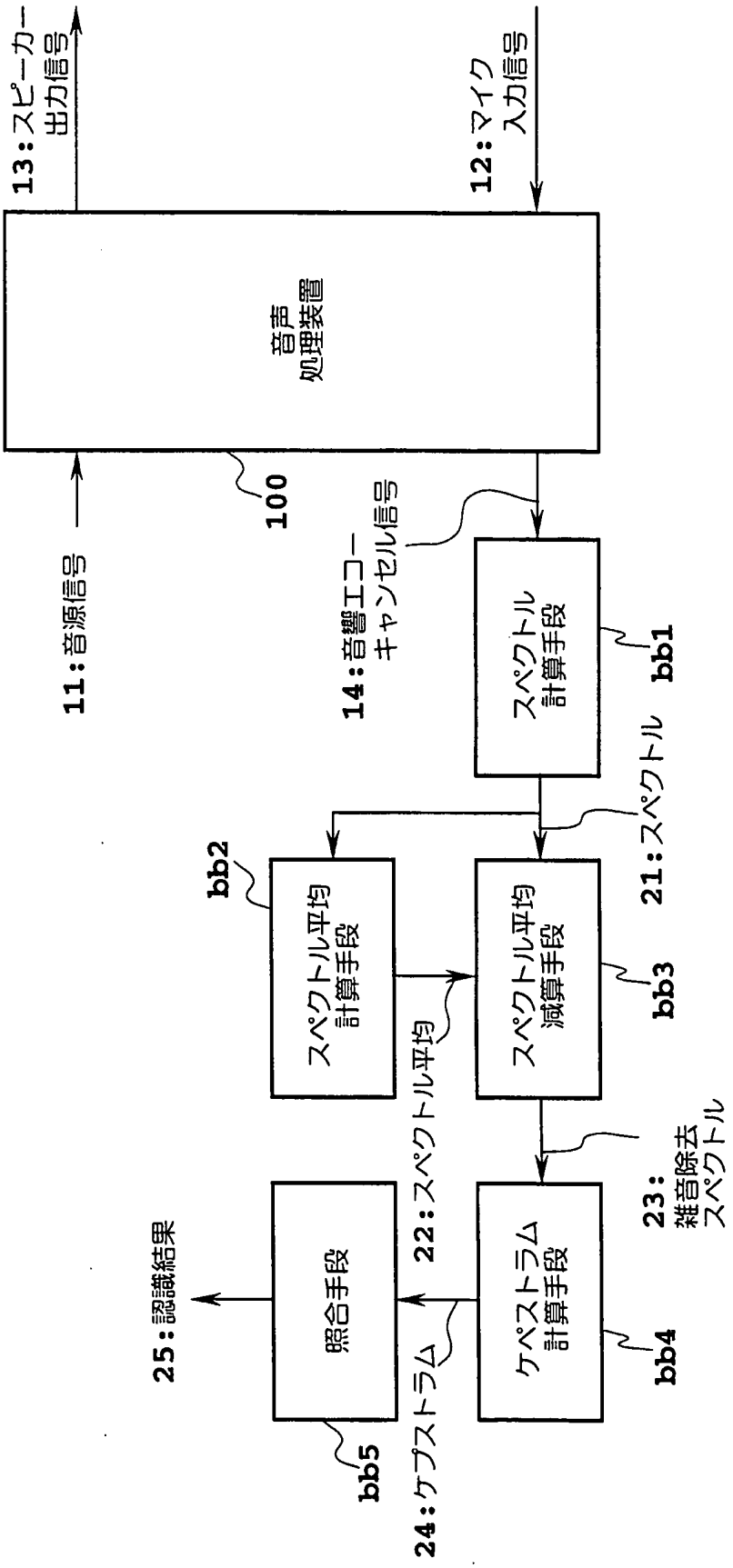

13：スピーカー出力信号
12：マイク入力信号
音声処理装置
100
11：音源信号
14：音響エコーキャンセル信号
スペクトル計算手段
bb1
21：スペクトル
bb2
スペクトル平均計算手段
スペクトル減算手段
bb3
22：スペクトル平均
23：雑音除去スペクトル
25：認識結果
照合手段
ケプストラム計算手段
bb4
24：ケプストラム
bb5

FIG.16

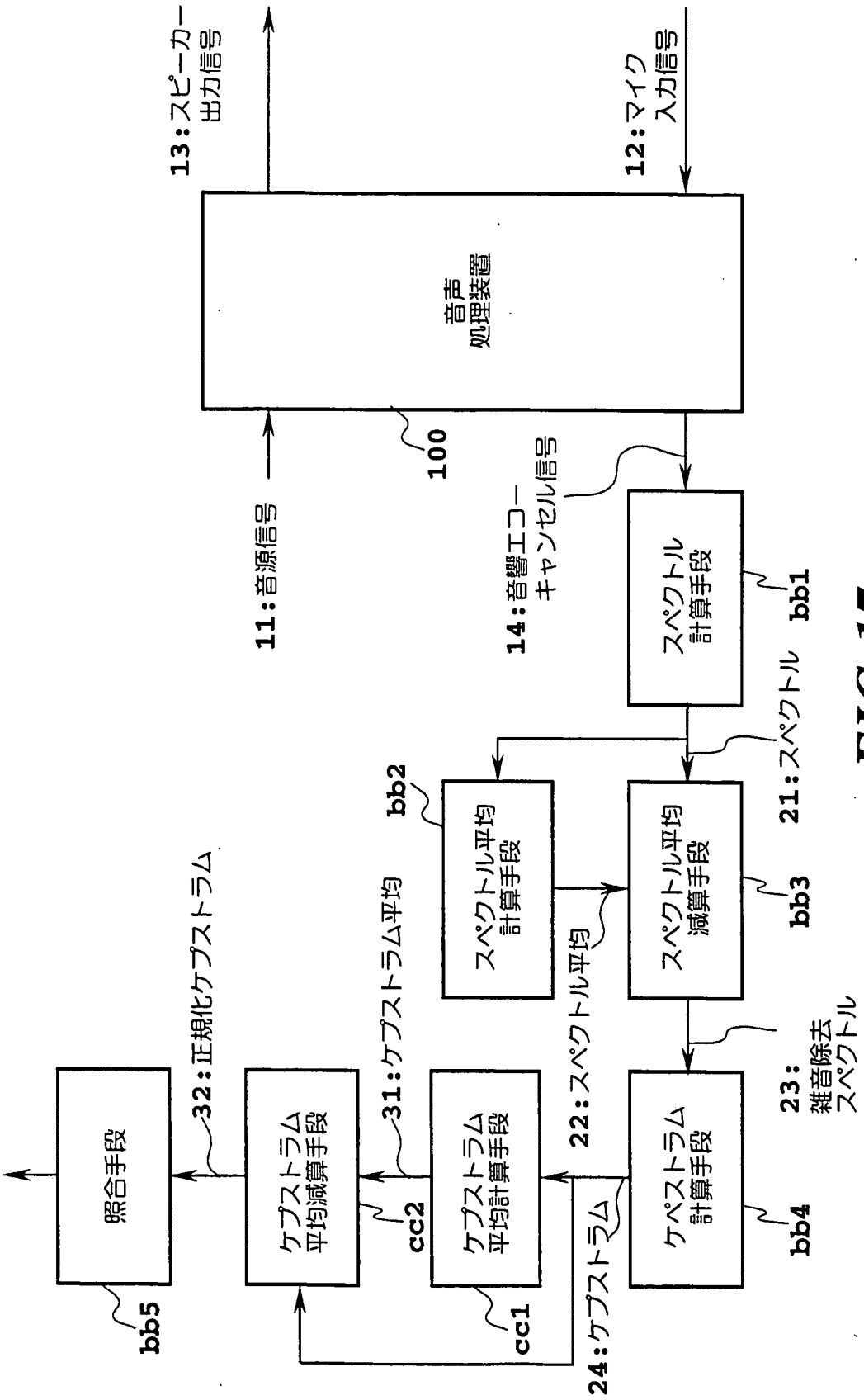

13:スピーカー出力信号
12:マイク入力信号
音声処理装置
100
11:音源信号
14:音響エコーキャンセル信号
スペクトル計算手段
bb1
21:スペクトル
スペクトル平均計算手段
bb2
スペクトル平均減算手段
bb3
31:ケプストラム平均
22:スペクトル平均
23:雑音除去スペクトル
ケプストラム計算手段
bb4
32:正規化ケプストラム
ケプストラム平均減算手段
cc2
ケプストラム平均計算手段
cc1
24:ケプストラム
照合手段
bb5

FIG.17

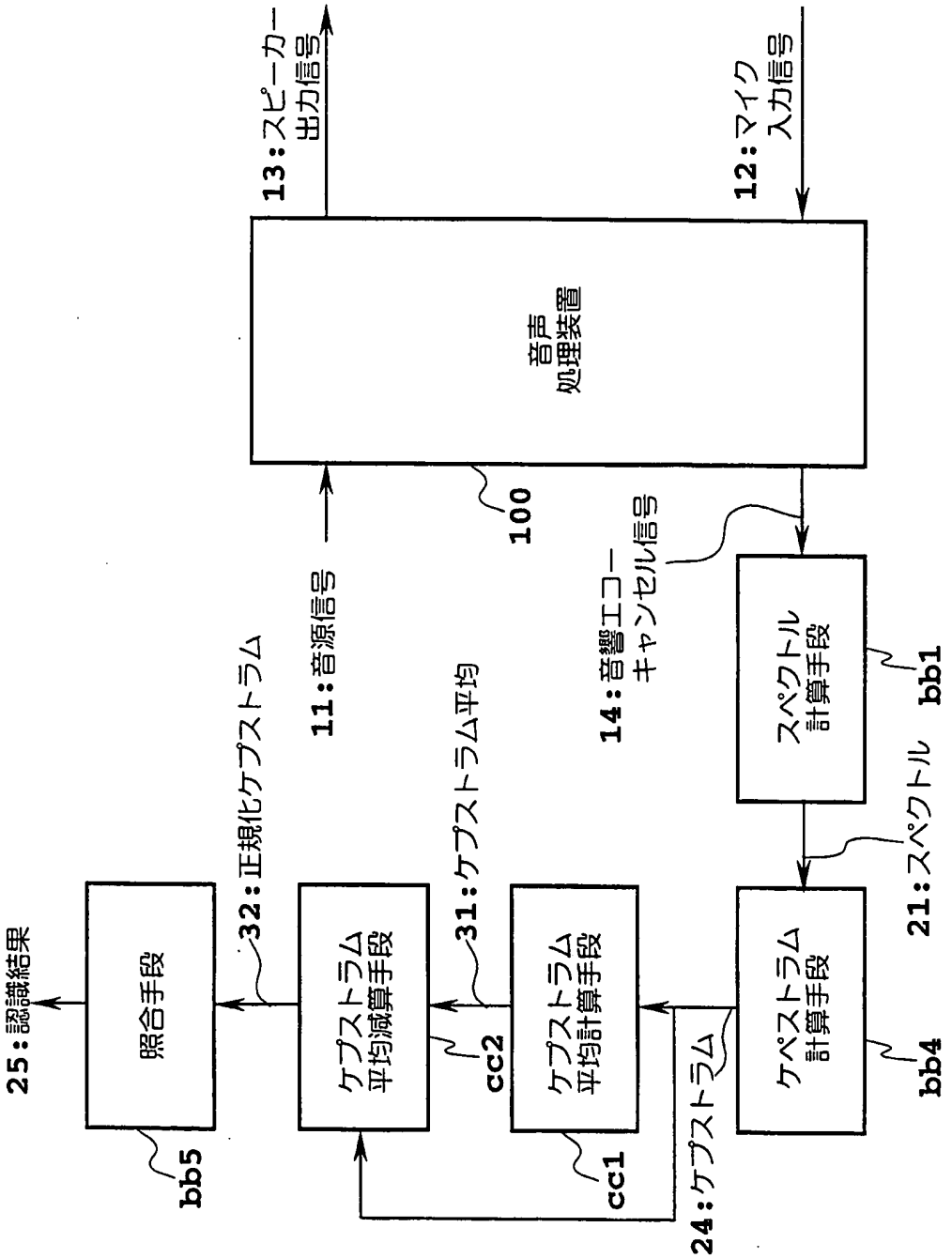

13:スピーカー出力信号
12:マイク入力信号
音声処理装置
100
11:音源信号
14:音響エコーキャンセル信号
スペクトル計算手段
bb1
21:スペクトル
31:ケプストラム平均
32:正規化ケプストラム
25:認識結果
照合手段
bb5
ケプストラム平均減算手段
cc2
ケプストラム平均計算手段
cc1
24:ケプストラム
ケプストラム計算手段
bb4

FIG.18

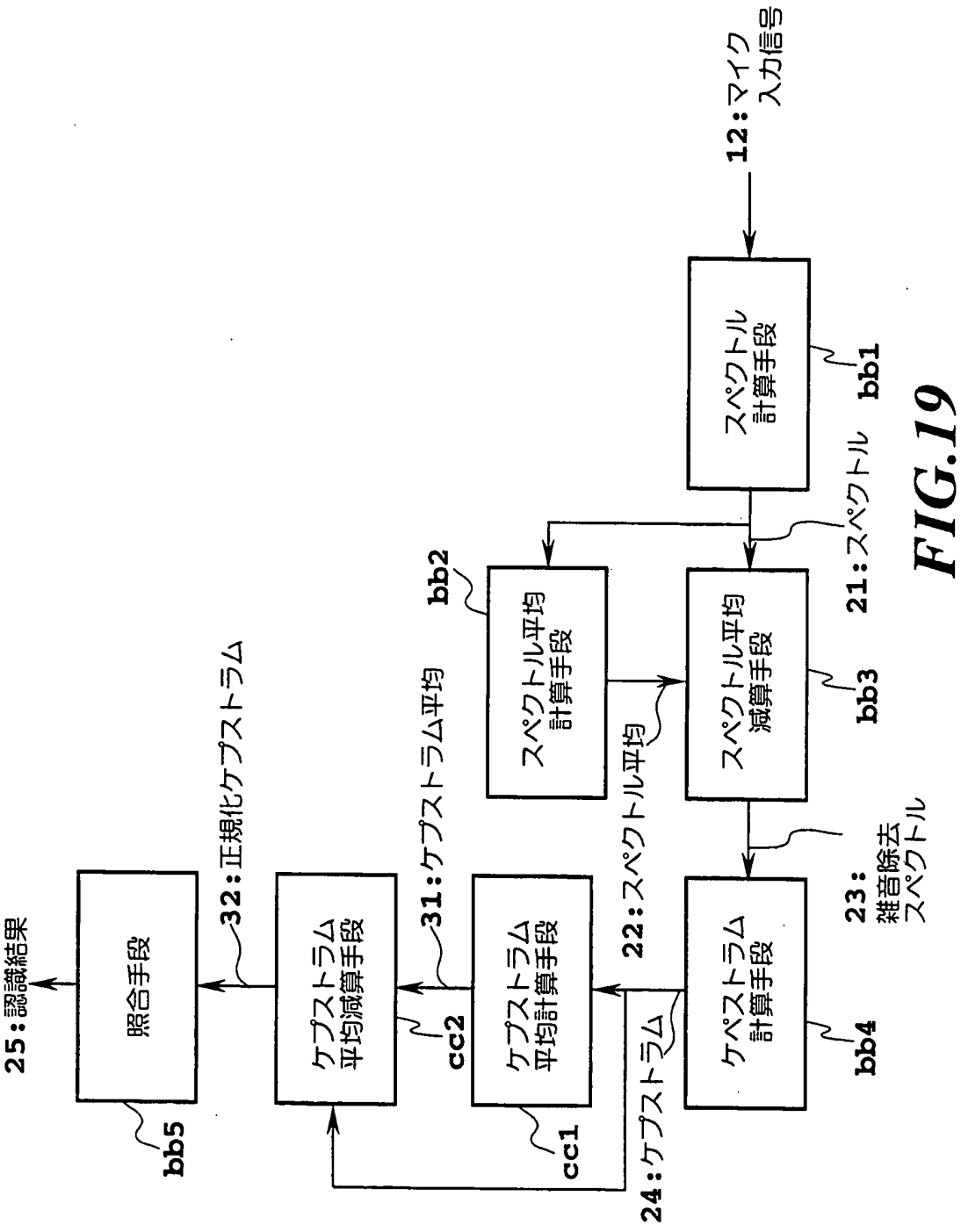

12:マイク入力信号
スペクトル計算手段
bb1
21:スペクトル
スペクトル平均減算手段
bb3
スペクトル平均計算手段
bb2
22:スペクトル平均
23:雑音除去スペクトル
ケプストラム計算手段
bb4
24:ケプストラム
ケプストラム平均計算手段
cc1
31:ケプストラム平均
ケプストラム平均減算手段
cc2
32:正規化ケプストラム
照合手段
bb5
25:認識結果

FIG.19

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/00915

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl^6 H04R3/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl^6 H04R3/02, H03H17/00, H04M1/16, H04B3/23

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926-1995
Kokai Jitsuyo Shinan Koho 1971-1995

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP, A, 63-18797 (Matsushita Electric Industrial Co., Ltd.), January 26, 1988 (26. 01. 88) (Family: none)	1, 7
Y	JP, A, 7-66757 (Nippon Telegraph and Telephone Corp.), March 10, 1995 (10. 03. 95) (Family: none)	2, 3, 8, 9

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
June 2, 1998 (02. 06. 98)	June 16, 1998 (16. 06. 98)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

国際調査報告

国際出願番号　PCT／JP98／00915

A.　発明の属する分野の分類（国際特許分類（IPC））
Int. cl⁶ H04R 3／02

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））
Int. cl⁶ H04R 3／02　H03H 17／00　H04M 1／16　H04B 3／23

最小限資料以外の資料で調査を行った分野に含まれるもの
日本国実用新案公法　1926－1995
日本国公開実用新案公法　1971－1995

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X	JP, A, 63－18797, (松下電器産業株式会社), 26. 1月. 1988 (26. 01. 88), ファミリなし	1, 7
Y	JP, A, 7－66757, (日本電信電話株式会社), 10. 3月. 1995 (10. 03. 95), ファミリなし	2, 3, 8, 9

☐ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」先行文献ではあるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
02. 06. 98	16.06.98

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）	5H	7525
日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	新宮佳典 印		
	電話番号　03－3581－1101　内線　3530		

様式PCT／ISA／210（第2ページ）（1992年7月）